PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2002-063209**
(43)Date of publication of application : **28.02.2002**

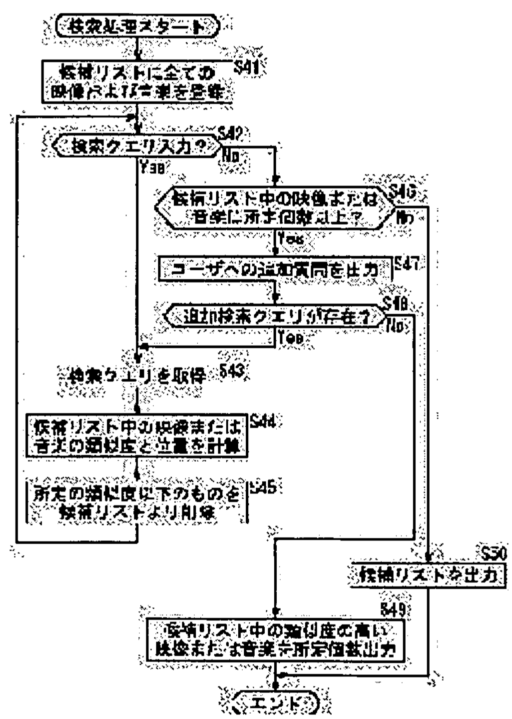
(51)Int.Cl. G06F 17/30
G06F 17/60

(21)Application number : **2000-250530**
(22)Date of filing : **22.08.2000**

(71)Applicant : **SONY CORP**
(72)Inventor : **ABE MOTOTSUGU**
NISHIGUCHI MASAYUKI
AKAGIRI KENZO

(54) INFORMATION PROCESSOR, ITS METHOD, INFORMATION SYSTEM, AND RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To retrieve the contents of an ambiguous impression by an interactive format.
SOLUTION: A retrieving processor judges whether a retrieving query is inputted or not in a step S42, and at the time of judging the input of the retrieving query, acquires the inputted retrieving queries, calculates the similarity of the queries, and, deletes the queries whose similarity is not more than a prescribed value from a candidate list in steps S43 to S45. At the time of judging no input of the retrieving query, the processor judges the number of contents in the candidate list is not less than a prescribed number or not in a step S46, and when the number of contents is not less than the prescribed number, outputs an additional query to a user in a step S47.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]An information processor when it has the following and said questionnaire entries are shown, wherein said calculating means computes further similarity of said contents registered into said candidate list based on said search condition of an addition inputted from a device besides the above.
Holding mechanism holding a candidate list which registered contents.
A calculating means which computes similarity of said contents registered into said candidate list based on a search condition inputted from other devices.
A deleting means which deletes said contents to which it judges, and said similarity computed by said calculating means corresponds [whether it is larger than a predetermined threshold and] when said similarity is smaller than a predetermined threshold from said candidate list.
A presenting means which presents questionnaire entries to a device besides the above when a total of said contents which remain in said candidate list is more than the predetermined number, as a result of being deleted by said deleting means.

[Claim 2]A transmitting means which transmits said candidate list to a device besides the above when there are few totals of said contents which remain in said candidate list than the predetermined number, as a result of being deleted by said deleting means, The information processor according to claim 1 having further a distribution means which distributes said contents to other devices when a demand of offer of said contents registered into said candidate list transmitted by said transmitting means from a device besides the above is received.
[Claim 3]The information processor according to claim 2, wherein it has further an acquisition means which acquires User Information of a device besides the above, and an authentication means which attests said User Information acquired by said acquisition means and said distribution means distributes said contents based on an authentication result by said authentication means.
[Claim 4]The information processor according to claim 1 having further a recording device which records said similarity computed by said calculating means and a similar position in said contents on said candidate list.
[Claim 5]The information processor according to claim 1, wherein said contents are picture image data or music data.
[Claim 6]The information processor according to claim 1, wherein form of said search condition contains a text about a text about an image, and music, an image, a sound, singing voice, humming, or music.
[Claim 7]The contents of said search condition A title name, a performer's name, a composer name, a songwriter name, Information relevant to a conductor name, a genre, words, a musical piece, humming or a performance by singing voice, information relevant to a musical piece, words, an actor name, an image, a reproduction image, and an image, or the information

processor according to claim 1 containing those parts.

[Claim 8]A maintenance step holding a candidate list which registered contents, and a calculation step which computes similarity of said contents registered into said candidate list based on a search condition inputted from other devices, A deletion step which deletes said contents to which it judges, and said similarity computed by processing of said calculation step corresponds [whether it is larger than a predetermined threshold and] when said similarity is smaller than a predetermined threshold from said candidate list, As a result of being deleted by processing of said deletion step, when a total of said contents which remain in said candidate list is more than the predetermined number, a device besides the above is received, An information processing method when said questionnaire entries are shown including a presentation step which presents questionnaire entries, wherein said calculation step computes further similarity of said contents registered into said candidate list based on said search condition of an addition inputted from a device besides the above.

[Claim 9]A maintenance step holding a candidate list which registered contents, and a calculation step which computes similarity of said contents registered into said candidate list based on a search condition inputted from other devices, A deletion step which deletes said contents to which it judges, and said similarity computed by processing of said calculation step corresponds [whether it is larger than a predetermined threshold and] when said similarity is smaller than a predetermined threshold from said candidate list, As a result of being deleted by processing of said deletion step, when a total of said contents which remain in said candidate list is more than the predetermined number, a device besides the above is received, When said questionnaire entries are shown including a presentation step which presents questionnaire entries, said calculation step, A recording medium with which a program which a computer computing further similarity of said contents registered into said candidate list based on said search condition of an addition inputted from a device besides the above can read is recorded.

[Claim 10]An information processing system which consists of the 1st information processor and 2nd information processor, comprising:

Holding mechanism holding a candidate list in which said 1st information processor registered contents.

A calculating means which computes similarity of said contents registered into said candidate list based on a search condition inputted from said 2nd information processor.

A deleting means which deletes said contents to which it judges, and said similarity computed by said calculating means corresponds [whether it is larger than a predetermined threshold and] when said similarity is smaller than a predetermined threshold from said candidate list.

As a result of being deleted by said deleting means, when a total of said contents which remain in said candidate list is more than the predetermined number, said 2nd information processor is received, The 1st transmitting means that transmits said search condition for said 2nd information processor to search said contents to said 1st information processor including a presenting means which presents questionnaire entries.

A reception means which receives said questionnaire entries shown from said 1st information processor.

The 2nd transmitting means that transmits said additional search condition to the 1st information processor when replying to said questionnaire entries received by said reception means.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the information processor and the method, information processing system, and recording medium which enabled it to search with interactive mode the contents in the vague impression which a user demands especially about an information processor and a method, an information processing system, and a recording medium, for example.
[0002]
[Description of the Prior Art]These days, various electronic commerce technology has come to be performed with the spread of the network systems represented by the Internet. For example, electronic commerce technology is performed by the method of a merchandise purchase person choosing goods from the commodity catalogs published by the homepage etc., and purchasing, or carrying out the direct entry of the trade name, and purchasing it, when the merchandise purchase person knows the already purchased trade name.
[0003]Thus, it is an effective system when the merchandise purchase person grasps the goods to purchase exactly in electronic commerce technology.
[0004]
[Problem(s) to be Solved by the Invention]However, in the case of an image, music (contents), etc., goods a buyer, The vague impression of contents to purchase, some scenes, a part of melody, some words, some words, Or the preview, the clip for advertisement, etc. are only memorized (record), and contents cannot be expressed in a concrete form like a contents name (title name), a performer's name, or a composer name in many cases.
[0005]By the way, in the case of the commercial transaction in the conventional shop front, the buyer can check a trade name by explaining a vague impression to a salesperson by a trade name's becoming clear or performing an audition and a preview at a shop front in the process in which it explains. That is, in the case of the conventional commercial transaction, the buyer can

purchase goods based on ambiguous (it is fuzzy) information.
[0006]However, in the case of electronic commerce technology, the buyer had a technical problem which cannot purchase contents based on ambiguous information.
[0007]This invention is made in view of such a situation, and enables it to search with interactive mode the contents in the vague impression which a user demands.
[0008]
[Means for Solving the Problem]Holding mechanism holding a candidate list in which an information processor of this invention registered contents, A calculating means which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deleting means which judges whether similarity computed by calculating means is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by a deleting means, when a total of contents which remain in a candidate list is more than the predetermined number, it has a presenting means which presents questionnaire entries to other devices and questionnaire entries are shown, A calculating means computes further similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0009]A transmitting means which transmits a candidate list to other devices when it has few totals of contents which remain in a candidate list than the predetermined number, as a result of an information processor of this invention being deleted by a deleting means, When a demand of offer of contents registered into a candidate list transmitted by transmitting means from other devices is received, a distribution means which distributes contents to other devices can be established further.
[0010]The information processor of this invention can establish further an acquisition means which acquires User Information of other devices, and an authentication means which attests User Information acquired by acquisition means, and the distribution means can distribute contents based on an authentication result by an authentication means.
[0011]The information processor of this invention can establish further a recording device which records similarity computed by calculating means and a similar position in contents on a candidate list.
[0012]Contents can be used as picture image data or music data.
[0013]The form of a search condition can contain a text about a text about an image, and music, an image, a sound, singing voice, humming, or music.
[0014]The contents of the search condition can contain information relevant to a title name, a performer's name, a composer name, a songwriter name, a conductor name, a genre, words, a musical piece, humming or a performance by singing voice, information relevant to a musical piece, words, an actor name, an image, a reproduction image, and an image, or those parts.
[0015]A maintenance step holding a candidate list in which an information processing method of this invention registered contents, A calculation step which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deletion step which judges whether similarity computed by processing of a calculation step is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by processing of a deletion step, when a total of contents which remain in a candidate list is more than the predetermined number and questionnaire entries are shown including a presentation step which presents questionnaire entries to other devices, A calculation step computes further

similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0016]A program currently recorded on a recording medium of this invention, A maintenance step holding a candidate list which registered contents, and a calculation step which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deletion step which judges whether similarity computed by processing of a calculation step is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by processing of a deletion step, when a total of contents which remain in a candidate list is more than the predetermined number and questionnaire entries are shown including a presentation step which presents questionnaire entries to other devices, A calculation step computes further similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0017]In an information processor of this invention, an information processing method, and a program currently recorded on a recording medium, Similarity of contents registered into a candidate list based on a search condition inputted from other devices is computed, and when computed similarity is smaller than a predetermined threshold, corresponding contents are deleted from a candidate list. And when a total of contents which remain in a candidate list is more than the predetermined number, questionnaire entries are shown to other devices and similarity of contents is further computed based on a search condition of an addition inputted from other devices.
[0018]Holding mechanism in which this invention holds a candidate list in which the 1st information processor registered contents, A calculating means which computes similarity of contents registered into a candidate list based on a search condition inputted from the 2nd information processor, A deleting means which judges whether similarity computed by calculating means is larger than a predetermined threshold, and deletes said corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by a deleting means, when a total of contents which remain in a candidate list is more than the predetermined number, the 2nd information processor is received, In order that the 2nd information processor may search contents including a presenting means which presents questionnaire entries, it is characterized by an information processing system comprising the following.
The 1st transmitting means that transmits a search condition to the 1st information processor.
A reception means which receives questionnaire entries shown from the 1st information processor.
The 2nd transmitting means that transmits an additional search condition to the 1st information processor when replying to questionnaire entries received by a reception means.

[0019]Based on a search condition which is the 1st information processor and is inputted from the 2nd information processor in an information processing system of this invention, When similarity of contents registered into a candidate list has similarity smaller than a predetermined threshold computed and computed, When a total of contents which remain in a candidate list is more than the predetermined number, to the 2nd information processor, corresponding contents are deleted from a candidate list, they are presented by questionnaire entries, and with the 2nd information processor. A search condition for searching contents is transmitted to the 1st information processor, and when replying to questionnaire entries shown from the 1st

information processor, an additional search condition is transmitted to the 1st information processor.
[0020]
[Embodiment of the Invention]Drawing 1 shows the example of composition of the search system which applied this invention. In this search system, it comprises the server system 1 connected via the Internet 2 and the terminal unit 3-1 thru/or 3-n (hereafter, when these terminal units 3-1 thru/or 3-n do not need to be distinguished separately, the terminal unit 3 is only called).
[0021]The server system 1 is a system which comprises two or more computers, and performs content retrieval processing mentioned later based on a server program and a CGI (Common Gateway Interface) script. The server system 1 charges a charge of search, a charge of distribution, etc. of contents to the terminal unit 3 again.
[0022]The terminal unit 3 is a computer and performs the WWW (World Wide Web) browser memorized by the hard disk drive (HDD) 29 which the CPU21 (drawing 3) builds in. The WWW browser performed with the terminal unit 3 by accessing the homepage which the server system 1 opens based on a user's instructions, From the server system 1, the HTML (Hyper TextMarkup Language) file transmitted via the Internet 2 is received, and the picture corresponding to the HTML file is displayed on the outputting part 27 (drawing 3).
[0023]Drawing 2 is a block diagram showing the detailed example of composition of a server system.
[0024]While the front-end processor 11 outputs the retrieval query (keyword of a broad sense used for search) transmitted from the terminal unit 3 via the Internet 2 to the retrieval server 12, The search results outputted from the retrieval server 12 are outputted to the terminal unit 3 via the Internet 2. Here, a retrieval query is the text about the music or the image for which it wishes, a sound, singing voice, humming, music, an image, or a picture.
[0025]The front-end processor 11 distributes the predetermined contents read based on the demand to the terminal unit 3 while notifying again the audition (preview) of contents or the demand of purchase transmitted from the terminal unit 3 to an image / music server 13. The front-end processor 11 outputs the accounting information outputted from the fee collection server 14 to the terminal unit 3 while notifying further User Information transmitted from the terminal unit 3 to the fee collection server 14.
[0026]The retrieval server 12 searches contents based on the retrieval query inputted from the front-end processor 11. At this time, the retrieval server 12 outputs questionnaire entries to the front-end processor 11 if needed.
[0027]All the images and music are accumulated in the image / music server 13. An image / music server 13 reads a predetermined image or music based on the audition (preview) of contents or the demand of purchase notified from the front-end processor 11.
[0028]The fee collection server 14 performs fee collection to the terminal unit 3 based on User Information notified from the front-end processor 11.
[0029]Drawing 3 is a block diagram showing the detailed example of composition of the terminal unit 3. Although a graphic display is omitted, the front-end processor 11 and the retrieval server 12 which were mentioned above, the image / music server 13, or the fee collection server 14 is constituted similarly.
[0030]CPU(Central Processing Unit) 21 performs various kinds of processings according to the program memorized by ROM(Read Only Memory) 22 and the hard disk drive 29. In RAM(Random Access Memory) 23, CPU21 performs various kinds of processings, and also a

required program and data are suitably memorized. CPU21, ROM22, and RAM23 are connected also to I/O interface 25 while being mutually connected via the bus 24.
[0031]In I/O interface 25, a keyboard, a ten key, a mouse, The input part 26, LCD (LiquidCrystalDisplay) which comprise a microphone, a digital camera, etc., The outputting part 27 which comprises CRT (Cathode Ray Tube), a loudspeaker, etc., the communications department 28 which communicates with the Internet 2, and the hard disk drive 29 are connected. The drive 30 for installing a program is connected to I/O interface 25 if needed, and it is equipped with the magnetic disk 41, the optical disc 42, the magneto-optical disc 43, or the semiconductor memory 44.
[0032]Drawing 4 is a block diagram showing the detailed example of composition of the retrieval server 12.
[0033]To the retrieval query of the text (information expressed in written form) inputted from the front-end processor 11, the text-processing part 51 performs predetermined processing, and outputs it to the retrieving processor 56. After the text-processing part 51 separates two or more retrieval queries inputted simultaneously from the front-end processor 11, it generates an image / music characteristic quantity, and, specifically, outputs it to the retrieving processor 56. The image / music characteristic quantity generated here are the texts itself.
[0034]To the retrieval query of the sound (information expressed with a user's own sound) inputted from the front-end processor 11, the voice processing part 52 performs predetermined processing, and outputs it to the retrieving processor 56. When the inputted sound or singing voice is changed into a text using speech recognition technology and two or more retrieval queries are inputted, after the voice processing part 52 separates it, it generates an image / music characteristic quantity, and, specifically, outputs it to the retrieving processor 56. The image / music characteristic quantity generated here are the texts itself.
[0035]The details of speech recognition technology are proposed by "the Furui:sound and voice engineering, Kindai Kagaku Sha, and 1992", for example.
[0036]To the retrieval query of the music (for example, information showing the music performed by FM broadcasting etc. (broadcast)) inputted from the front-end processor 11, the music treating part 53 performs predetermined processing, and outputs it to the retrieving processor 56. Specifically, the music treating part 53 extracts the characteristic quantity of the inputted music using music analytical skills. The music characteristic quantity generated here is a text in which the genres (a lock, a classic, etc.) of digital data, such as an output swing of a band pass filter (BPF), or music are shown, for example.
[0037]The details of the extraction method of music characteristic quantity are indicated by "Kenyon:Signal Recoginition system and method, US Patent 5210820", for example, The details of the method of discriminating the genre of a musical piece from music for example, "Han:Genre Classification systemof TV sound Signals Based on a Spectrogram Analysis,IEEE Trans.on Consumer Electronics, Vol. . It is proposed by 44 and No.1-1998."
[0038]The singing voice into which the singing voice humming treating part 54 was inputted from the front-end processor 11 (information expressed with a user's sound own [with melody and words]), Or to the retrieval query of humming (information expressed with a user's own sound who does not have words although it has melody), predetermined processing is performed and it outputs to the retrieving processor 56. The singing voice humming treating part 54 was not performed by the original player of the musical piece, and, specifically, extracts the characteristic quantity showing the melody of a musical piece from the performance mainly reproduced by the user itself using a feature extraction method. The music characteristic quantity generated here is

the digital data showing the height and length of a note, for example, is expressed with MIDI form.

[0039]The details of the extraction method of humming characteristic quantity are proposed by "the music retrieval system and database system using the Kosugi:humming, 119-9, Information Processing Society of Japan, and 1999", for example.

[0040]To the retrieval query of the image (information expressed by the animation) inputted from the front-end processor 11, or a picture (information expressed with the still picture), the graphic processing part 55 performs predetermined processing, and outputs it to the retrieving processor 56. Specifically, the graphic processing part 55 extracts characteristic quantity, such as an image by which television broadcasting was carried out, the clip image and 1 frame image of an image which were inputted from the front-end processor 11 and which were recorded, or a user's own sketch drawing picture, using the extraction method of image characteristic quantity. The image characteristic quantity generated here is a color histogram of an image, a border line, or a motion vector, and is expressed as digital data, for example.

[0041]The details of the extraction method of image characteristic quantity are proposed by "the automatic indexing method in the Nagasaka:color video picture, object heuristics, the Information Processing Society of Japan paper magazine, Vol.33, No.4, pp.543-50-1992", for example.

[0042]All the images by which the retrieving processor 56 is accumulated in the retrieving database 57 based on the image / music characteristic quantity inputted, respectively from the text-processing part 51 thru/or the graphic processing part 55, or musical characteristic quantity, Similarity R_{xy} between the inputted image / music characteristic quantity is computed according to a following formula (1).

R_{xy}=(number of characters in agreement)/(the length of a retrieval query) ... (1)

[0043]The above-mentioned formula (1) is used when computing similarity from the characteristic quantity of text format. When computing similarity from the characteristic quantity of digital data form, it is computed according to a following formula (2).

$R_{xy}=(x\text{-}y)/\mathrm{root}\ (|x|^2|y|^2)$ ... (2)

Here, x expresses the characteristic quantity of the inputted image or music, and y expresses the characteristic quantity of the image accumulated in the retrieving database 57, or music.

[0044]Based on similarity R_{xy} computed by the above-mentioned formula (1) or the formula (2), the retrieving processor 56 detects the music ($R_{xy}=1$) which is thoroughly in agreement, the music which has a similar portion, or an image ($0<R_{xy}<1$), and outputs it as a candidate list mentioned later.

[0045]The retrieving database 57 is constituted by memory storage, such as a hard disk drive and a magneto-optical disc, and the control processor which controls it. By the characteristic quantity of the image and music which are used for search being beforehand registered into the retrieving database 57, for example, using database languages, such as SQL (Structured Query Language), it collects as data record of one or a small number, and is managed.

[0046]Drawing 5 shows the example of the retrieval query inputted into the text-processing part 51 thru/or the graphic processing part 55 from the terminal unit 3.

[0047]As for the 1st entry, in the case of the example of drawing 5, the contents express the title name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 2nd entry, the contents express the performer's name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. The 3rd entry expresses the retrieval query which the sex of the player of a musical piece or a national origin name, and form

become from a text or a sound in the contents. As for the 4th entry, the contents express the composer name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 5th entry, the contents express the songwriter name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 6th entry, the contents express the conductor name of a musical piece, or its retrieval query which form becomes from a text or a sound in part.

[0048]The 7th entry expresses the retrieval query which the genre of a musical piece and form become from a text, a sound, or music in the contents. The 8th entry expresses the retrieval query which some of words of a musical piece or words, and form become from a text, a sound, or singing voice in the contents. The 9th entry expresses the retrieval query which a part of musical piece or musical piece, and form that the contents were recorded become from music. The 10th entry expresses humming or a performance according [the contents] to singing voice, and its retrieval query which form becomes from singing voice or humming in part. The 11th entry expresses the retrieval query which the information relevant to [in addition to this] a musical piece in the contents, including a composition fiscal year, a sale fiscal year, etc., and form become from a text or a sound. As for the 12th entry, the contents express the title name of an image, or its retrieval query which form becomes from a text or a sound in part.

[0049]As for the 13th entry, the contents express further again the implementor name of an image, or its retrieval query which form becomes from a text or a sound in part. The 14th entry expresses the retrieval query which some of words or words, and form that the contents are included in an image become from a text or a sound. The 15th entry expresses the actor name of the main actor for whom the contents are included in an image, or its retrieval query which form becomes from a text or a sound in part. The 16th entry expresses the retrieval query which a part of image or image, and form that the contents were recorded become from an image or a picture. The 17th entry expresses the retrieval query which the image and the picture, and form that the contents imitated or reproduced an image or its part become from an image or a picture. The 18th entry expresses the retrieval query which the information relevant to [in addition to this] an image in the contents, including a manufacture fiscal year, an open fiscal year, etc., and form become from a text or a sound.

[0050]As shown also in the form of the retrieval query shown in <u>drawing 5</u>, to the text-processing part 51 and the voice processing part 52. The retrieval query entered by the retrieval query entered by the 1st thru/or the 8th, the 11th, or the 15th or the retrieval query entered by the 18th is inputted, respectively. The retrieval query entered by the 7th or the 9th is inputted into the music treating part 53. The retrieval query entered by the 10th is inputted into the singing voice humming treating part 54. The retrieval query entered by the 16th or the 17th is inputted into the graphic processing part 55.

[0051]<u>Drawing 6</u> shows the example of the candidate list outputted from the retrieving processor 56.

[0052]In the case of the example of <u>drawing 6</u>, in similarity, 97%, a title expresses a moon libber and, as for the 1st entry, the query position expresses the contents for 3 minutes and 24 seconds (it is hereafter indicated as 3'24). In the 2nd entry, 88%, a title expresses The Umbrellas of Cherbourg and the query position expresses [similarity] the contents of 1'20. If a title sings the 3rd entry to rain 83% in similarity, the query position expresses the contents of 2'30. In the 4th entry, 77%, a title expresses Somewhere Over the Rainbow and the query position expresses [similarity] the contents of 0'05.

[0053]Here, query positions are the retrieval query inputted by the user into the image registered

into the retrieving database 57, or music, and a similar position. For example, in the case of the contents entered by the 1st, a thing similar to the retrieval query inputted by the user exists in the position (time) which carried out 24 second passage for 3 minutes from the head (0 minute and 00 seconds) of the musical piece of a moon libber, and the similarity is 97%. In the retrieval processing mentioned later, this query position is used, when trying listening or holding a preview.

[0054]When a retrieval query cannot pinpoint the position in an image or music like a title name, For example, the position (portions of the scene used against a title background and the rust of a musical piece) generally in the image or music represented is described to the candidate list as a default value.

[0055]Next, with reference to the flow chart of drawing 7 and drawing 8, the message distribution processing of contents (an image or music) which the front-end processor 11 of the server system 1 performs is explained.

[0056]In Step S1, the front-end processor 11 judges whether it is accessed from the terminal unit 3 via the Internet 2, and it stands by until it is accessed from the terminal unit 3. And in Step S1, if accessed from the terminal unit 3, it will progress to Step S2 and the front-end processor 11 will distribute the HTML file memorized by the hard disk drive to build in to the terminal unit 3 via the Internet 2. Thereby, an initial entry screen as shown in drawing 9 is displayed on the outputting part 27 of the terminal unit 3.

[0057]In the case of the example of drawing 9, the retrieval query input area 71 is displayed on an initial entry screen. The input part 26 is used by the user of the terminal unit 3, a retrieval query is inputted into the retrieval query input area 71, and a retrieval query is inputted to the server system 1 by pushing the search start button 72. Even if a user inputs a sound, singing voice, or humming using a microphone or he not only inputs retrieval queries, such as a text, but it inputs an image or a picture using a digital camera, for example, he can do it.

[0058]To Step S3, it returns to drawing 7, sets [are,], and the front-end processor 11 acquires the retrieval query transmitted from the terminal unit 3 via the Internet 2. In step S4, the front-end processor 11 transmits the retrieval query acquired by processing of Step S3 to the retrieval server 12. Thereby, the retrieval server 12 performs retrieval processing mentioned later based on the retrieval query supplied from the front-end processor 11, and outputs search results.

[0059]In Step S5, the front-end processor 11 acquires the output from the retrieval server 12. In Step S6, the front-end processor 11 judges whether it is that the output of the retrieval server 12 acquired by processing of Step S5 is a question to a user, and when it judges with it being a question to a user, it progresses to Step S7.

[0060]In Step S7, the front-end processor 11 distributes the HTML file about the questionnaire entries outputted from the retrieval server 12 to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 10 is displayed on the outputting part 27 of the terminal unit 3.

[0061]In the case of the example of drawing 10, the question to a user and the answer input area 81 of the terminal unit 3 are displayed. When a reply (additional retrieval query) is inputted into the answer input area 81 by the user who checked these questionnaire entries and OK button 82 is pushed, the reply to that question is transmitted to the server system 1.

[0062]Returning to drawing 7, in Step S8, the front-end processor 11 acquires the reply (additional retrieval query) transmitted from the terminal unit 3 via the Internet 2, and repeats the processing returned and mentioned above to step S4.

[0063]In Step S6, when the output of the retrieval server 12 acquired by processing of Step S5 judges with it not being a question to a user, it progresses to step S9 and the front-end processor

11 judges further whether it is that the output of the retrieval server 12 is a candidate list. In step S9, when judged with the output of the retrieval server 12 not being a candidate list, the HTML file about search failure is distributed to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 11 is displayed on the outputting part 27 of the terminal unit 3.
[0064]In the case of the example of drawing 11, message "search went wrong. It indicates that there is no applicable candidate." When OK button 91 is pushed by the user who checked this message, it can return to the initial entry screen shown in drawing 9.
[0065]In step S9, when it judges that the output of the retrieval server 12 is a candidate list, it progresses to Step S11 and the front-end processor 11 distributes the HTML file about a candidate list to the terminal unit 3 via the Internet 2. Thereby, the screen of a candidate list as shown in drawing 12 is displayed on the outputting part 27 of the terminal unit 3.
[0066]In the case of the example of drawing 12, a contents name and its similarity are displayed on the high order of similarity. The selection button 101-1 of either of the contents displayed on the candidate list thru/or 101-4 are chosen by the user of the terminal unit 3, An audition / preview of predetermined contents, or purchase is required from the server system 1 by pushing either audition/preview button 102 or the buy button 103. When the end button 104 is pushed, it can return to the initial entry screen shown in drawing 9.
[0067]Returning to drawing 7, in Step S12, the front-end processor 11 acquires the user input (an audition/preview, purchase, or end) transmitted from the terminal unit 3 via the Internet 2.
[0068]In Step S13, the front-end processor 11 judges whether it is that the user input from the terminal unit 3 acquired by processing of Step S12 is an audition or a preview, and when it judges with their being an audition or a preview, it progresses to Step S14. In Step S14, the front-end processor 11 determines an audition portion or a preview portion based on the query position of the candidate list shown in drawing 6. That is, in the retrieval processing mentioned later, since a query position is described by the candidate list, the predetermined section including the position is determined as an audition or a preview portion.
[0069]For example, in the example of drawing 6, when the audition of the moon libber entered by the 1st is required, the front-end processor 11 determines the predetermined section as an audition portion from the query position for 3 minutes and 24 seconds of the contents. Since the portion which the user has imagined can be used for an audition or a preview by this, a user is a short time and can perform the check of the contents of contents effectively.
[0070]In Step S15, the front-end processor 11 transmits the audition portion or preview portion determined by processing of Step S14 to an image / music server 13. Thereby, an image / music server 13 reads the predetermined audition portion or preview portion of contents based on the audition portion or preview portion supplied from the front-end processor 11.
[0071]In Step S16, the front-end processor 11 acquires the contents of the audition portion read from the image / music server 13, or a preview portion. In Step S17, the front-end processor 11 provides the terminal unit 3 with the contents of the audition portion acquired by processing of Step S16, or a preview portion via the Internet 2 (transmission). Thereby, a screen as shown in drawing 13 is displayed on the outputting part 27 of the terminal unit 3.
[0072]In the case of the example of drawing 13, the contents of an audition portion or a preview portion are reproduced (output). When the button 111 of "repeating once again" is pushed by the user who tried listening them or previewed contents, the contents of an audition portion or a preview portion are reproduced again. It can return to the screen of the candidate list shown in drawing 12 by pushing the end button 112.
[0073]Return to drawing 7 and the user input from the terminal unit 3 acquired by processing of

Step S12 in Step S13, When judged with their not being an audition or a preview, it progresses to Step S18 and the front-end processor 11 judges further whether it is that the user input from the terminal unit 3 is purchase.

[0074]Here, when the buy button 103 shown in drawing 12 is pushed, a screen as shown in drawing 14 is further displayed on the outputting part 27 of the terminal unit 3. While the message "input User Information" is displayed to the user of the terminal unit 3 in the case of the example of drawing 14, the user ID input area 121 and the password input area 122 are displayed. User ID is inputted into the user ID input area 121 by the user of the terminal unit 3, the password of user ID is entered into the password input area 122, and User Information is inputted to the server system 1 by pushing OK button 123. User ID is a number of a credit card, a number of a cellular phone, etc. which a user owns, for example.

[0075]It returns to drawing 8, and in Step S18, when judged with a user input being purchase, it progresses to Step S19 and the front-end processor 11 acquires User Information transmitted from the terminal unit 3 via the Internet 2. In Step S20, the front-end processor 11 transmits User Information acquired by processing of Step S19 to the fee collection server 14. Thereby, the fee collection server 14 performs accounting mentioned later based on User Information supplied from the front-end processor 11, and outputs a processing result.

[0076]In Step S21, the front-end processor 11 acquires the output from the fee collection server 14. When it judges with the front-end processor 11 judging whether it is that the output of the fee collection server 14 acquired by processing of Step S21 is "permission", and the output of the fee collection server 14 being "permission" in Step S22, it progresses to Step S23.

[0077]In Step S23, the front-end processor 11 notifies an image / music server 13 that the output of contents was permitted. Thereby, an image / music server 13 reads the predetermined contents sold in response to the notice of an output permission from the front-end processor 11.

[0078]In Step S24, the front-end processor 11 acquires the contents read by the image / music server 13. In Step S25, the front-end processor 11 distributes the contents acquired by processing of Step S24 to the terminal unit 3 via the Internet 2.

[0079]In Step S22, the output of the fee collection server 14 acquired by processing of Step S21, When it was not "permission", i.e., judged with the output of the fee collection server 14 being "disapproval", it progresses to Step S26 and the front-end processor 11 distributes the HTML file about "disapproval" to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 15 is displayed on the outputting part 27 of the terminal unit 3.

[0080]In the case of the example of drawing 15, the message "contents are not downloadable" is displayed to the user of the terminal unit 3.

[0081]Thus, in the message distribution processing of contents, the contents to search can be narrowed down by repeating a question to the retrieval query which the user inputted if needed.

[0082]Next, with reference to the flow chart of drawing 16, the retrieval processing which the retrieving processor 56 of the retrieval server 12 performs is explained.

[0083]In Step S41, the retrieving processor 56 registers into a candidate list the characteristic quantity of all the images registered into the retrieving database 57, and music. In Step S42, when it judges whether it is that the image / music characteristic quantity (retrieval query) generated in the text-processing part 51 thru/or the graphic processing part 55 were inputted and judges with the retrieval query having been inputted, he follows the retrieving processor 56 to Step S43.

[0084]In Step S43, the retrieving processor 56 acquires the retrieval query inputted by processing of Step S42. In Step S44, the retrieving processor 56 computes similarity R_{xy} between the

retrieval query (an image / music characteristic quantity) acquired by processing of Step S43, and all the images in a candidate list or musical characteristic quantity according to the above-mentioned formula (1) or a formula (2).
[0085]In Step S45, the retrieving processor 56 deletes the following [predetermined similarity] from a candidate list among similarity R_{xy} computed by processing of Step S44, and repeats the processing returned and mentioned above to Step S42. The threshold of the similarity deleted from a candidate list is set up arbitrarily.
[0086]In Step S42, when it judges with the retrieval query not being inputted, it progresses to Step S46 and the front-end processor 11 judges whether it is that the image or music in a candidate list is more than a prescribed number (for example, ten pieces). In Step S46, when judged with the image or music in a candidate list being more than a prescribed number, it progresses to Step S47 and the retrieving processor 56 outputs the additional questions to the user of the terminal unit 3 to the front-end processor 11.
[0087]Thereby, the front-end processor 11 distributes the additional questions supplied from the retrieving processor 56 to the terminal unit 3 via the Internet 2, and displays a screen as shown in drawing 10 on the outputting part 27 of the terminal unit 3. When a reply (additional retrieval query) is inputted into the answer input area 81 by the user who checked this screen and OK button 82 is pushed, the reply to that question is transmitted to the server system 1.
[0088]In Step S48, the retrieving processor 56 repeats the processing after it which he followed to Step S43 and was mentioned above, when it judges whether it is that the additional retrieval query was inputted and judges with the additional retrieval query having been inputted.
[0089]In Step S48, when judged with the additional retrieval query not being inputted, it progresses to Step S49 and the retrieving processor 56 outputs the high image or music of a prescribed number (for example, ten pieces) of similarity in a candidate list to the front-end processor 11. The number of the image outputted to the front-end processor 11 or music is set up arbitrarily.
[0090]In Step S46, when judged with the image or music in a candidate list not being more than a prescribed number (for example, ten pieces), it progresses to Step S50, the retrieving processor 56 outputs a candidate list to the front-end processor 11, and processing is ended.
[0091]Thus, in retrieval processing, the contents to search can be narrowed down by repeating a question to a user until the contents in a candidate list are settled within a prescribed number. When there is no reply (additional retrieval query) to a question, the contents of the high prescribed number of similarity can be made into search results.
[0092]Next, with reference to the flow chart of drawing 17, the accounting which the fee collection server 14 performs is explained. As for this processing, in Yes (a user output is purchase), the decision processing of Step S18 of drawing 8 is started.
[0093]In Step S61, the fee collection server 14 receives User Information transmitted from the front-end processor 11, and acquires the user ID contained in the User Information. In Step S62, based on the user ID acquired by processing of Step S61, the user pays the fee collection server 14 to the network management contractor who does not illustrate, and it asks whether to be possible (settlement of accounts is possible).
[0094]In Step S63, when the fee collection server 14 acquired the reply from a network management contractor, and it judges whether it is that payment is possible, it pays and it judges with it being possible, it progresses to Step S64 and it outputs "permission" to the front-end processor 11. In Step S63, when the user of the terminal unit 3 pays and it is judged with it not being possible, it progresses to Step S65, the fee collection server 14 outputs "disapproval" to the

front-end processor 11, and processing is ended.
[0095]Thus, in accounting, a check of the person himself/herself and the method of paying can be determined based on the user ID acquired from the front-end processor 11. Payment methods include a credit card transaction or the vicarious execution settlement of accounts by a network operator, for example.
[0096]Although it was made to perform retrieval processing above via the Internet 2, this invention may be made to perform retrieval processing via the wire communication not only using this but cable TV etc., or the radio using a terrestrial wave or satellite broadcasting waves. In the case of radio, a portable telephone, PDA (Personal Digital Assistant), etc. may be sufficient as the terminal unit 3.
[0097]As mentioned above, the server system 1 can extract a search condition by repeating some questions to the ambiguous information demanded by the user. Therefore, an effect as taken below is acquired by using this invention.
(1) In an electronic video distribution system or an electronic music distribution system, an image or music can be searched from the vague impression which cannot necessarily carry out [keyword]-izing.
(2) The user can try listening it or preview an image or music to purchase before purchase.
(3) A user becomes possible [choosing and purchasing desired goods by interactive check] from a vague image.
[0098]Although a series of processings mentioned above can also be performed by hardware, they can also be performed by software. The computer by which the program which constitutes the software is included in hardware for exclusive use when performing a series of processings by software, Or it is installed in the personal computer etc. which can perform various kinds of functions, for example, are general-purpose, etc. from a recording medium by installing various kinds of programs.
[0099]. As shown in drawing 3, this recording medium is distributed apart from a computer in order to provide a user with a program. The magnetic disk 41 (a floppy disk is included) with which the program is recorded, the optical disc 42 (CD-ROM (Compact Disk-Read Only Memory).) . DVD (Digital Versatile Disk) is included. It is not only constituted by the package media which consist of the magneto-optical disc 43 (MD (Mini-Disk) is included) or the semiconductor memory 44, but, It comprises ROM22 with which a user is provided in the state where it was beforehand included in the computer and on which the program is recorded, the hard disk drive 29, etc.
[0100]In this specification, even if the processing serially performed in accordance with an order that the step which describes the program recorded on a recording medium was indicated is not of course necessarily processed serially, it also includes a parallel target or the processing performed individually.
[0101]In this specification, a system expresses the whole device constituted by two or more devices.
[0102]
[Effect of the Invention]As mentioned above, according to the program currently recorded on the information processor of this invention, the method, and the recording medium. The similarity of the contents registered into the candidate list based on the search condition inputted from other devices is computed, and when the computed similarity is smaller than a predetermined threshold, corresponding contents are deleted from a candidate list. And other devices are received when the total of the contents which remain in the candidate list is more than the

predetermined number, Questionnaire entries are shown, and since the similarity of contents was further computed based on the search condition of the addition inputted from other devices, the contents in a vague impression can be searched with interactive mode.
[0103]According to the information processing system of this invention, the 1st information processor, Based on the search condition inputted from the 2nd information processor, the similarity of the contents registered into the candidate list is computed, Delete the contents which correspond when the computed similarity is smaller than a predetermined threshold from a candidate list, and when the total of the contents which remain in the candidate list is more than the predetermined number, the 2nd information processor is received, When replying to the questionnaire entries which presented questionnaire entries, and the 2nd information processor transmitted the search condition for searching contents to the 1st information processor, and were shown from the 1st information processor, Since the additional search condition was transmitted to the 1st information processor, the contents in a vague impression can be searched with interactive mode.

[Translation done.]

TECHNICAL FIELD

[Field of the Invention]This invention relates to the information processor and the method, information processing system, and recording medium which enabled it to search with interactive mode the contents in the vague impression which a user demands especially about an information processor and a method, an information processing system, and a recording medium, for example.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]These days, various electronic commerce technology has come to be performed with the spread of the network systems represented by the Internet. For example, electronic commerce technology is performed by the method of a merchandise purchase person choosing goods from the commodity catalogs published by the homepage etc., and purchasing, or carrying out the direct entry of the trade name, and purchasing it, when the merchandise purchase person knows the already purchased trade name.
[0003]Thus, it is an effective system when the merchandise purchase person grasps the goods to purchase exactly in electronic commerce technology.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]As mentioned above, according to the program currently recorded on the information processor of this invention, the method, and the recording medium. The similarity of the contents registered into the candidate list based on the search condition inputted from other devices is computed, and when the computed similarity is smaller than a predetermined threshold, corresponding contents are deleted from a candidate list. And other devices are received when the total of the contents which remain in the candidate list is more than the predetermined number, Questionnaire entries are shown, and since the similarity of contents was further computed based on the search condition of the addition inputted from other devices, the contents in a vague impression can be searched with interactive mode.

[0103]According to the information processing system of this invention, the 1st information processor, Based on the search condition inputted from the 2nd information processor, the similarity of the contents registered into the candidate list is computed, Delete the contents which correspond when the computed similarity is smaller than a predetermined threshold from a candidate list, and when the total of the contents which remain in the candidate list is more than the predetermined number, the 2nd information processor is received, When replying to the questionnaire entries which presented questionnaire entries, and the 2nd information processor transmitted the search condition for searching contents to the 1st information processor, and were shown from the 1st information processor, Since the additional search condition was transmitted to the 1st information processor, the contents in a vague impression can be searched with interactive mode.

[Translation done.]

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]However, in the case of an image, music (contents), etc., goods a buyer, The vague impression of contents to purchase, some scenes, a part of melody, some words, some words, Or the preview, the clip for advertisement, etc. are only memorized (record), and contents cannot be expressed in a concrete form like a contents name (title name), a performer's name, or a composer name in many cases.
[0005]By the way, in the case of the commercial transaction in the conventional shop front, the buyer can check a trade name by explaining a vague impression to a salesperson by a trade name's becoming clear or performing an audition and a preview at a shop front in the process in which it explains. That is, in the case of the conventional commercial transaction, the buyer can purchase goods based on ambiguous (it is fuzzy) information.
[0006]However, in the case of electronic commerce technology, the buyer had a technical problem which cannot purchase contents based on ambiguous information.
[0007]This invention is made in view of such a situation, and enables it to search with interactive mode the contents in the vague impression which a user demands.

[Translation done.]

* NOTICES *

`JPO and INPIT are not responsible for any damages caused by the use of this translation.`

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

MEANS

[Means for Solving the Problem]Holding mechanism holding a candidate list in which an information processor of this invention registered contents, A calculating means which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deleting means which judges whether similarity computed by calculating means is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by a deleting means, when a total of contents which remain in a candidate list is more than the predetermined number, it has a presenting means which presents questionnaire entries to other devices and questionnaire entries are shown, A calculating means computes further similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0009]A transmitting means which transmits a candidate list to other devices when it has few totals of contents which remain in a candidate list than the predetermined number, as a result of an information processor of this invention being deleted by a deleting means, When a demand of offer of contents registered into a candidate list transmitted by transmitting means from other devices is received, a distribution means which distributes contents to other devices can be established further.
[0010]The information processor of this invention can establish further an acquisition means which acquires User Information of other devices, and an authentication means which attests User Information acquired by acquisition means, and the distribution means can distribute contents based on an authentication result by an authentication means.
[0011]The information processor of this invention can establish further a recording device which records similarity computed by calculating means and a similar position in contents on a candidate list.
[0012]Contents can be used as picture image data or music data.
[0013]The form of a search condition can contain a text about a text about an image, and music, an image, a sound, singing voice, humming, or music.
[0014]The contents of the search condition can contain information relevant to a title name, a performer's name, a composer name, a songwriter name, a conductor name, a genre, words, a

musical piece, humming or a performance by singing voice, information relevant to a musical piece, words, an actor name, an image, a reproduction image, and an image, or those parts.
[0015]A maintenance step holding a candidate list in which an information processing method of this invention registered contents, A calculation step which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deletion step which judges whether similarity computed by processing of a calculation step is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by processing of a deletion step, when a total of contents which remain in a candidate list is more than the predetermined number and questionnaire entries are shown including a presentation step which presents questionnaire entries to other devices, A calculation step computes further similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0016]A program currently recorded on a recording medium of this invention, A maintenance step holding a candidate list which registered contents, and a calculation step which computes similarity of contents registered into a candidate list based on a search condition inputted from other devices, A deletion step which judges whether similarity computed by processing of a calculation step is larger than a predetermined threshold, and deletes corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by processing of a deletion step, when a total of contents which remain in a candidate list is more than the predetermined number and questionnaire entries are shown including a presentation step which presents questionnaire entries to other devices, A calculation step computes further similarity of contents registered into a candidate list based on a search condition of an addition inputted from other devices.
[0017]In an information processor of this invention, an information processing method, and a program currently recorded on a recording medium, Similarity of contents registered into a candidate list based on a search condition inputted from other devices is computed, and when computed similarity is smaller than a predetermined threshold, corresponding contents are deleted from a candidate list. And when a total of contents which remain in a candidate list is more than the predetermined number, questionnaire entries are shown to other devices and similarity of contents is further computed based on a search condition of an addition inputted from other devices.
[0018]Holding mechanism in which this invention holds a candidate list in which the 1st information processor registered contents, A calculating means which computes similarity of contents registered into a candidate list based on a search condition inputted from the 2nd information processor, A deleting means which judges whether similarity computed by calculating means is larger than a predetermined threshold, and deletes said corresponding contents from a candidate list when similarity is smaller than a predetermined threshold, As a result of being deleted by a deleting means, when a total of contents which remain in a candidate list is more than the predetermined number, the 2nd information processor is received, In order that the 2nd information processor may search contents including a presenting means which presents questionnaire entries, it is characterized by an information processing system comprising the following.
The 1st transmitting means that transmits a search condition to the 1st information processor.
A reception means which receives questionnaire entries shown from the 1st information processor.

The 2nd transmitting means that transmits an additional search condition to the 1st information processor when replying to questionnaire entries received by a reception means.

[0019]Based on a search condition which is the 1st information processor and is inputted from the 2nd information processor in an information processing system of this invention, When similarity of contents registered into a candidate list has similarity smaller than a predetermined threshold computed and computed, When a total of contents which remain in a candidate list is more than the predetermined number, to the 2nd information processor, corresponding contents are deleted from a candidate list, they are presented by questionnaire entries, and with the 2nd information processor. A search condition for searching contents is transmitted to the 1st information processor, and when replying to questionnaire entries shown from the 1st information processor, an additional search condition is transmitted to the 1st information processor.
[0020]
[Embodiment of the Invention]Drawing 1 shows the example of composition of the search system which applied this invention. In this search system, it comprises the server system 1 connected via the Internet 2 and the terminal unit 3-1 thru/or 3-n (hereafter, when these terminal units 3-1 thru/or 3-n do not need to be distinguished separately, the terminal unit 3 is only called).
[0021]The server system 1 is a system which comprises two or more computers, and performs content retrieval processing mentioned later based on a server program and a CGI (Common Gateway Interface) script. The server system 1 charges a charge of search, a charge of distribution, etc. of contents to the terminal unit 3 again.
[0022]The terminal unit 3 is a computer and performs the WWW (World Wide Web) browser memorized by the hard disk drive (HDD) 29 which the CPU21 (drawing 3) builds in. The WWW browser performed with the terminal unit 3 by accessing the homepage which the server system 1 opens based on a user's instructions, From the server system 1, the HTML (Hyper TextMarkup Language) file transmitted via the Internet 2 is received, and the picture corresponding to the HTML file is displayed on the outputting part 27 (drawing 3).
[0023]Drawing 2 is a block diagram showing the detailed example of composition of a server system.
[0024]While the front-end processor 11 outputs the retrieval query (keyword of a broad sense used for search) transmitted from the terminal unit 3 via the Internet 2 to the retrieval server 12, The search results outputted from the retrieval server 12 are outputted to the terminal unit 3 via the Internet 2. Here, a retrieval query is the text about the music or the image for which it wishes, a sound, singing voice, humming, music, an image, or a picture.
[0025]The front-end processor 11 distributes the predetermined contents read based on the demand to the terminal unit 3 while notifying again the audition (preview) of contents or the demand of purchase transmitted from the terminal unit 3 to an image / music server 13. The front-end processor 11 outputs the accounting information outputted from the fee collection server 14 to the terminal unit 3 while notifying further User Information transmitted from the terminal unit 3 to the fee collection server 14.
[0026]The retrieval server 12 searches contents based on the retrieval query inputted from the front-end processor 11. At this time, the retrieval server 12 outputs questionnaire entries to the front-end processor 11 if needed.
[0027]All the images and music are accumulated in the image / music server 13. An image /

music server 13 reads a predetermined image or music based on the audition (preview) of contents or the demand of purchase notified from the front-end processor 11.
[0028]The fee collection server 14 performs fee collection to the terminal unit 3 based on User Information notified from the front-end processor 11.
[0029]Drawing 3 is a block diagram showing the detailed example of composition of the terminal unit 3. Although a graphic display is omitted, the front-end processor 11 and the retrieval server 12 which were mentioned above, the image / music server 13, or the fee collection server 14 is constituted similarly.
[0030]CPU(Central Processing Unit) 21 performs various kinds of processings according to the program memorized by ROM(Read Only Memory) 22 and the hard disk drive 29. In RAM(Random Access Memory) 23, CPU21 performs various kinds of processings, and also a required program and data are suitably memorized. CPU21, ROM22, and RAM23 are connected also to I/O interface 25 while being mutually connected via the bus 24.
[0031]In I/O interface 25, a keyboard, a ten key, a mouse, The input part 26, LCD (LiquidCrystalDisplay) which comprise a microphone, a digital camera, etc., The outputting part 27 which comprises CRT (Cathode Ray Tube), a loudspeaker, etc., the communications department 28 which communicates with the Internet 2, and the hard disk drive 29 are connected. The drive 30 for installing a program is connected to I/O interface 25 if needed, and it is equipped with the magnetic disk 41, the optical disc 42, the magneto-optical disc 43, or the semiconductor memory 44.
[0032]Drawing 4 is a block diagram showing the detailed example of composition of the retrieval server 12.
[0033]To the retrieval query of the text (information expressed in written form) inputted from the front-end processor 11, the text-processing part 51 performs predetermined processing, and outputs it to the retrieving processor 56. After the text-processing part 51 separates two or more retrieval queries inputted simultaneously from the front-end processor 11, it generates an image / music characteristic quantity, and, specifically, outputs it to the retrieving processor 56. The image / music characteristic quantity generated here are the texts itself.
[0034]To the retrieval query of the sound (information expressed with a user's own sound) inputted from the front-end processor 11, the voice processing part 52 performs predetermined processing, and outputs it to the retrieving processor 56. When the inputted sound or singing voice is changed into a text using speech recognition technology and two or more retrieval queries are inputted, after the voice processing part 52 separates it, it generates an image / music characteristic quantity, and, specifically, outputs it to the retrieving processor 56. The image / music characteristic quantity generated here are the texts itself.
[0035]The details of speech recognition technology are proposed by "the Furui:sound and voice engineering, Kindai Kagaku Sha, and 1992", for example.
[0036]To the retrieval query of the music (for example, information showing the music performed by FM broadcasting etc. (broadcast)) inputted from the front-end processor 11, the music treating part 53 performs predetermined processing, and outputs it to the retrieving processor 56. Specifically, the music treating part 53 extracts the characteristic quantity of the inputted music using music analytical skills. The music characteristic quantity generated here is a text in which the genres (a lock, a classic, etc.) of digital data, such as an output swing of a band pass filter (BPF), or music are shown, for example.
[0037]The details of the extraction method of music characteristic quantity are indicated by "Kenyon:Signal Recoginition system and method, US Patent 5210820", for example, The details

of the method of discriminating the genre of a musical piece from music for example, "Han:Genre Classification system of TV sound Signals Based on a Spectrogram Analysis,IEEE Trans.on Consumer Electronics,. It is proposed by Vol.44 and No.1-1998."
[0038]The singing voice into which the singing voice humming treating part 54 was inputted from the front-end processor 11 (information expressed with a user's sound own [with melody and words]), Or to the retrieval query of humming (information expressed with a user's own sound who does not have words although it has melody), predetermined processing is performed and it outputs to the retrieving processor 56. The singing voice humming treating part 54 was not performed by the original player of the musical piece, and, specifically, extracts the characteristic quantity showing the melody of a musical piece from the performance mainly reproduced by the user itself using a feature extraction method. The music characteristic quantity generated here is the digital data showing the height and length of a note, for example, is expressed with MIDI form.
[0039]The details of the extraction method of humming characteristic quantity are proposed by "the music retrieval system and database system using the Kosugi:humming, 119-9, Information Processing Society of Japan, and 1999", for example.
[0040]To the retrieval query of the image (information expressed by the animation) inputted from the front-end processor 11, or a picture (information expressed with the still picture), the graphic processing part 55 performs predetermined processing, and outputs it to the retrieving processor 56. Specifically, the graphic processing part 55 extracts characteristic quantity, such as an image by which television broadcasting was carried out, the clip image and 1 frame image of an image which were inputted from the front-end processor 11 and which were recorded, or a user's own sketch drawing picture, using the extraction method of image characteristic quantity. The image characteristic quantity generated here is a color histogram of an image, a border line, or a motion vector, and is expressed as digital data, for example.
[0041]The details of the extraction method of image characteristic quantity are proposed by "the automatic indexing method in the Nagasaka:color video picture, object heuristics, the Information Processing Society of Japan paper magazine, Vol.33, No.4, pp.543-50-1992", for example.
[0042]All the images by which the retrieving processor 56 is accumulated in the retrieving database 57 based on the image / music characteristic quantity inputted, respectively from the text-processing part 51 thru/or the graphic processing part 55, or musical characteristic quantity, Similarity R_{xy} between the inputted image / music characteristic quantity is computed according to a following formula (1).
R_{xy}=(number of characters in agreement)/(the length of a retrieval query) ... (1)
[0043]The above-mentioned formula (1) is used when computing similarity from the characteristic quantity of text format. When computing similarity from the characteristic quantity of digital data form, it is computed according to a following formula (2).
$R_{xy}=(x\text{-}y)/\text{root}\ (|x|^2|y|^2)$ ... (2)
Here, x expresses the characteristic quantity of the inputted image or music, and y expresses the characteristic quantity of the image accumulated in the retrieving database 57, or music.
[0044]Based on similarity R_{xy} computed by the above-mentioned formula (1) or the formula (2), the retrieving processor 56 detects the music ($R_{xy}=1$) which is thoroughly in agreement, the music which has a similar portion, or an image ($0<R_{xy}<1$), and outputs it as a candidate list mentioned later.
[0045]The retrieving database 57 is constituted by memory storage, such as a hard disk drive and

a magneto-optical disc, and the control processor which controls it. By the characteristic quantity of the image and music which are used for search being beforehand registered into the retrieving database 57, for example, using database languages, such as SQL (Structured Query Language), it collects as data record of one or a small number, and is managed.
[0046]Drawing 5 shows the example of the retrieval query inputted into the text-processing part 51 thru/or the graphic processing part 55 from the terminal unit 3.
[0047]As for the 1st entry, in the case of the example of drawing 5, the contents express the title name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 2nd entry, the contents express the performer's name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. The 3rd entry expresses the retrieval query which the sex of the player of a musical piece or a national origin name, and form become from a text or a sound in the contents. As for the 4th entry, the contents express the composer name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 5th entry, the contents express the songwriter name of a musical piece, or its retrieval query which form becomes from a text or a sound in part. As for the 6th entry, the contents express the conductor name of a musical piece, or its retrieval query which form becomes from a text or a sound in part.
[0048]The 7th entry expresses the retrieval query which the genre of a musical piece and form become from a text, a sound, or music in the contents. The 8th entry expresses the retrieval query which some of words of a musical piece or words, and form become from a text, a sound, or singing voice in the contents. The 9th entry expresses the retrieval query which a part of musical piece or musical piece, and form that the contents were recorded become from music. The 10th entry expresses humming or a performance according [the contents] to singing voice, and its retrieval query which form becomes from singing voice or humming in part. The 11th entry expresses the retrieval query which the information relevant to [in addition to this] a musical piece in the contents, including a composition fiscal year, a sale fiscal year, etc., and form become from a text or a sound. As for the 12th entry, the contents express the title name of an image, or its retrieval query which form becomes from a text or a sound in part.
[0049]As for the 13th entry, the contents express further again the implementor name of an image, or its retrieval query which form becomes from a text or a sound in part. The 14th entry expresses the retrieval query which some of words or words, and form that the contents are included in an image become from a text or a sound. The 15th entry expresses the actor name of the main actor for whom the contents are included in an image, or its retrieval query which form becomes from a text or a sound in part. The 16th entry expresses the retrieval query which a part of image or image, and form that the contents were recorded become from an image or a picture. The 17th entry expresses the retrieval query which the image and the picture, and form that the contents imitated or reproduced an image or its part become from an image or a picture. The 18th entry expresses the retrieval query which the information relevant to [in addition to this] an image in the contents, including a manufacture fiscal year, an open fiscal year, etc., and form become from a text or a sound.
[0050]As shown also in the form of the retrieval query shown in drawing 5, to the text-processing part 51 and the voice processing part 52. The retrieval query entered by the retrieval query entered by the 1st thru/or the 8th, the 11th, or the 15th or the retrieval query entered by the 18th is inputted, respectively. The retrieval query entered by the 7th or the 9th is inputted into the music treating part 53. The retrieval query entered by the 10th is inputted into the singing voice humming treating part 54. The retrieval query entered by the 16th or the 17th is inputted into the

graphic processing part 55.
[0051]Drawing 6 shows the example of the candidate list outputted from the retrieving processor 56.
[0052]In the case of the example of drawing 6, in similarity, 97%, a title expresses a moon libber and, as for the 1st entry, the query position expresses the contents for 3 minutes and 24 seconds (it is hereafter indicated as 3'24). In the 2nd entry, 88%, a title expresses The Umbrellas of Cherbourg and the query position expresses [similarity] the contents of 1'20. If a title sings the 3rd entry to rain 83% in similarity, the query position expresses the contents of 2'30. In the 4th entry, 77%, a title expresses Somewhere Over the Rainbow and the query position expresses [similarity] the contents of 0'05.
[0053]Here, query positions are the retrieval query inputted by the user into the image registered into the retrieving database 57, or music, and a similar position. For example, in the case of the contents entered by the 1st, a thing similar to the retrieval query inputted by the user exists in the position (time) which carried out 24 second passage for 3 minutes from the head (0 minute and 00 seconds) of the musical piece of a moon libber, and the similarity is 97%. In the retrieval processing mentioned later, this query position is used, when trying listening or holding a preview.
[0054]When a retrieval query cannot pinpoint the position in an image or music like a title name, For example, the position (portions of the scene used against a title background and the rust of a musical piece) generally in the image or music represented is described to the candidate list as a default value.
[0055]Next, with reference to the flow chart of drawing 7 and drawing 8, the message distribution processing of contents (an image or music) which the front-end processor 11 of the server system 1 performs is explained.
[0056]In Step S1, the front-end processor 11 judges whether it is accessed from the terminal unit 3 via the Internet 2, and it stands by until it is accessed from the terminal unit 3. And in Step S1, if accessed from the terminal unit 3, it will progress to Step S2 and the front-end processor 11 will distribute the HTML file memorized by the hard disk drive to build in to the terminal unit 3 via the Internet 2. Thereby, an initial entry screen as shown in drawing 9 is displayed on the outputting part 27 of the terminal unit 3.
[0057]In the case of the example of drawing 9, the retrieval query input area 71 is displayed on an initial entry screen. The input part 26 is used by the user of the terminal unit 3, a retrieval query is inputted into the retrieval query input area 71, and a retrieval query is inputted to the server system 1 by pushing the search start button 72. Even if a user inputs a sound, singing voice, or humming using a microphone or he not only inputs retrieval queries, such as a text, but it inputs an image or a picture using a digital camera, for example, he can do it.
[0058]To Step S3, it returns to drawing 7, sets [are,], and the front-end processor 11 acquires the retrieval query transmitted from the terminal unit 3 via the Internet 2. In step S4, the front-end processor 11 transmits the retrieval query acquired by processing of Step S3 to the retrieval server 12. Thereby, the retrieval server 12 performs retrieval processing mentioned later based on the retrieval query supplied from the front-end processor 11, and outputs search results.
[0059]In Step S5, the front-end processor 11 acquires the output from the retrieval server 12. In Step S6, the front-end processor 11 judges whether it is that the output of the retrieval server 12 acquired by processing of Step S5 is a question to a user, and when it judges with it being a question to a user, it progresses to Step S7.
[0060]In Step S7, the front-end processor 11 distributes the HTML file about the questionnaire

entries outputted from the retrieval server 12 to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 10 is displayed on the outputting part 27 of the terminal unit 3.

[0061]In the case of the example of drawing 10, the question to a user and the answer input area 81 of the terminal unit 3 are displayed. When a reply (additional retrieval query) is inputted into the answer input area 81 by the user who checked these questionnaire entries and OK button 82 is pushed, the reply to that question is transmitted to the server system 1.

[0062]Returning to drawing 7, in Step S8, the front-end processor 11 acquires the reply (additional retrieval query) transmitted from the terminal unit 3 via the Internet 2, and repeats the processing returned and mentioned above to step S4.

[0063]In Step S6, when the output of the retrieval server 12 acquired by processing of Step S5 judges with it not being a question to a user, it progresses to step S9 and the front-end processor 11 judges further whether it is that the output of the retrieval server 12 is a candidate list. In step S9, when judged with the output of the retrieval server 12 not being a candidate list, the HTML file about search failure is distributed to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 11 is displayed on the outputting part 27 of the terminal unit 3.

[0064]In the case of the example of drawing 11, message "search went wrong. It indicates that there is no applicable candidate." When OK button 91 is pushed by the user who checked this message, it can return to the initial entry screen shown in drawing 9.

[0065]In step S9, when it judges that the output of the retrieval server 12 is a candidate list, it progresses to Step S11 and the front-end processor 11 distributes the HTML file about a candidate list to the terminal unit 3 via the Internet 2. Thereby, the screen of a candidate list as shown in drawing 12 is displayed on the outputting part 27 of the terminal unit 3.

[0066]In the case of the example of drawing 12, a contents name and its similarity are displayed on the high order of similarity. The selection button 101-1 of either of the contents displayed on the candidate list thru/or 101-4 are chosen by the user of the terminal unit 3, An audition / preview of predetermined contents, or purchase is required from the server system 1 by pushing either audition/preview button 102 or the buy button 103. When the end button 104 is pushed, it can return to the initial entry screen shown in drawing 9.

[0067]Returning to drawing 7, in Step S12, the front-end processor 11 acquires the user input (an audition/preview, purchase, or end) transmitted from the terminal unit 3 via the Internet 2.

[0068]In Step S13, the front-end processor 11 judges whether it is that the user input from the terminal unit 3 acquired by processing of Step S12 is an audition or a preview, and when it judges with their being an audition or a preview, it progresses to Step S14. In Step S14, the front-end processor 11 determines an audition portion or a preview portion based on the query position of the candidate list shown in drawing 6. That is, in the retrieval processing mentioned later, since a query position is described by the candidate list, the predetermined section including the position is determined as an audition or a preview portion.

[0069]For example, in the example of drawing 6, when the audition of the moon libber entered by the 1st is required, the front-end processor 11 determines the predetermined section as an audition portion from the query position for 3 minutes and 24 seconds of the contents. Since the portion which the user has imagined can be used for an audition or a preview by this, a user is a short time and can perform the check of the contents of contents effectively.

[0070]In Step S15, the front-end processor 11 transmits the audition portion or preview portion determined by processing of Step S14 to an image / music server 13. Thereby, an image / music server 13 reads the predetermined audition portion or preview portion of contents based on the audition portion or preview portion supplied from the front-end processor 11.

[0071]In Step S16, the front-end processor 11 acquires the contents of the audition portion read from the image / music server 13, or a preview portion. In Step S17, the front-end processor 11 provides the terminal unit 3 with the contents of the audition portion acquired by processing of Step S16, or a preview portion via the Internet 2 (transmission). Thereby, a screen as shown in drawing 13 is displayed on the outputting part 27 of the terminal unit 3.
[0072]In the case of the example of drawing 13, the contents of an audition portion or a preview portion are reproduced (output). When the button 111 of "repeating once again" is pushed by the user who tried listening them or previewed contents, the contents of an audition portion or a preview portion are reproduced again. It can return to the screen of the candidate list shown in drawing 12 by pushing the end button 112.
[0073]Return to drawing 7 and the user input from the terminal unit 3 acquired by processing of Step S12 in Step S13, When judged with their not being an audition or a preview, it progresses to Step S18 and the front-end processor 11 judges further whether it is that the user input from the terminal unit 3 is purchase.
[0074]Here, when the buy button 103 shown in drawing 12 is pushed, a screen as shown in drawing 14 is further displayed on the outputting part 27 of the terminal unit 3. While the message "input User Information" is displayed to the user of the terminal unit 3 in the case of the example of drawing 14, the user ID input area 121 and the password input area 122 are displayed. User ID is inputted into the user ID input area 121 by the user of the terminal unit 3, the password of user ID is entered into the password input area 122, and User Information is inputted to the server system 1 by pushing OK button 123. User ID is a number of a credit card, a number of a cellular phone, etc. which a user owns, for example.
[0075]It returns to drawing 8, and in Step S18, when judged with a user input being purchase, it progresses to Step S19 and the front-end processor 11 acquires User Information transmitted from the terminal unit 3 via the Internet 2. In Step S20, the front-end processor 11 transmits User Information acquired by processing of Step S19 to the fee collection server 14. Thereby, the fee collection server 14 performs accounting mentioned later based on User Information supplied from the front-end processor 11, and outputs a processing result.
[0076]In Step S21, the front-end processor 11 acquires the output from the fee collection server 14. When it judges with the front-end processor 11 judging whether it is that the output of the fee collection server 14 acquired by processing of Step S21 is "permission", and the output of the fee collection server 14 being "permission" in Step S22, it progresses to Step S23.
[0077]In Step S23, the front-end processor 11 notifies an image / music server 13 that the output of contents was permitted. Thereby, an image / music server 13 reads the predetermined contents sold in response to the notice of an output permission from the front-end processor 11.
[0078]In Step S24, the front-end processor 11 acquires the contents read by the image / music server 13. In Step S25, the front-end processor 11 distributes the contents acquired by processing of Step S24 to the terminal unit 3 via the Internet 2.
[0079]In Step S22, the output of the fee collection server 14 acquired by processing of Step S21, When it was not "permission", i.e., judged with the output of the fee collection server 14 being "disapproval", it progresses to Step S26 and the front-end processor 11 distributes the HTML file about "disapproval" to the terminal unit 3 via the Internet 2. Thereby, a screen as shown in drawing 15 is displayed on the outputting part 27 of the terminal unit 3.
[0080]In the case of the example of drawing 15, the message "contents are not downloadable" is displayed to the user of the terminal unit 3.
[0081]Thus, in the message distribution processing of contents, the contents to search can be

narrowed down by repeating a question to the retrieval query which the user inputted if needed.
[0082]Next, with reference to the flow chart of drawing 16, the retrieval processing which the retrieving processor 56 of the retrieval server 12 performs is explained.
[0083]In Step S41, the retrieving processor 56 registers into a candidate list the characteristic quantity of all the images registered into the retrieving database 57, and music. In Step S42, when it judges whether it is that the image / music characteristic quantity (retrieval query) generated in the text-processing part 51 thru/or the graphic processing part 55 were inputted and judges with the retrieval query having been inputted, he follows the retrieving processor 56 to Step S43.
[0084]In Step S43, the retrieving processor 56 acquires the retrieval query inputted by processing of Step S42. In Step S44, the retrieving processor 56 computes similarity R_{xy} between the retrieval query (an image / music characteristic quantity) acquired by processing of Step S43, and all the images in a candidate list or musical characteristic quantity according to the above-mentioned formula (1) or a formula (2).
[0085]In Step S45, the retrieving processor 56 deletes the following [predetermined similarity] from a candidate list among similarity R_{xy} computed by processing of Step S44, and repeats the processing returned and mentioned above to Step S42. The threshold of the similarity deleted from a candidate list is set up arbitrarily.
[0086]In Step S42, when it judges with the retrieval query not being inputted, it progresses to Step S46 and the front-end processor 11 judges whether it is that the image or music in a candidate list is more than a prescribed number (for example, ten pieces). In Step S46, when judged with the image or music in a candidate list being more than a prescribed number, it progresses to Step S47 and the retrieving processor 56 outputs the additional questions to the user of the terminal unit 3 to the front-end processor 11.
[0087]Thereby, the front-end processor 11 distributes the additional questions supplied from the retrieving processor 56 to the terminal unit 3 via the Internet 2, and displays a screen as shown in drawing 10 on the outputting part 27 of the terminal unit 3. When a reply (additional retrieval query) is inputted into the answer input area 81 by the user who checked this screen and OK button 82 is pushed, the reply to that question is transmitted to the server system 1.
[0088]In Step S48, the retrieving processor 56 repeats the processing after it which he followed to Step S43 and was mentioned above, when it judges whether it is that the additional retrieval query was inputted and judges with the additional retrieval query having been inputted.
[0089]In Step S48, when judged with the additional retrieval query not being inputted, it progresses to Step S49 and the retrieving processor 56 outputs the high image or music of a prescribed number (for example, ten pieces) of similarity in a candidate list to the front-end processor 11. The number of the image outputted to the front-end processor 11 or music is set up arbitrarily.
[0090]In Step S46, when judged with the image or music in a candidate list not being more than a prescribed number (for example, ten pieces), it progresses to Step S50, the retrieving processor 56 outputs a candidate list to the front-end processor 11, and processing is ended.
[0091]Thus, in retrieval processing, the contents to search can be narrowed down by repeating a question to a user until the contents in a candidate list are settled within a prescribed number. When there is no reply (additional retrieval query) to a question, the contents of the high prescribed number of similarity can be made into search results.
[0092]Next, with reference to the flow chart of drawing 17, the accounting which the fee collection server 14 performs is explained. As for this processing, in Yes (a user output is

purchase), the decision processing of Step S18 of drawing 8 is started.
[0093]In Step S61, the fee collection server 14 receives User Information transmitted from the front-end processor 11, and acquires the user ID contained in the User Information. In Step S62, based on the user ID acquired by processing of Step S61, the user pays the fee collection server 14 to the network management contractor who does not illustrate, and it asks whether to be possible (settlement of accounts is possible).
[0094]In Step S63, when the fee collection server 14 acquired the reply from a network management contractor, and it judges whether it is that payment is possible, it pays and it judges with it being possible, it progresses to Step S64 and it outputs "permission" to the front-end processor 11. In Step S63, when the user of the terminal unit 3 pays and it is judged with it not being possible, it progresses to Step S65, the fee collection server 14 outputs "disapproval" to the front-end processor 11, and processing is ended.
[0095]Thus, in accounting, a check of the person himself/herself and the method of paying can be determined based on the user ID acquired from the front-end processor 11. Payment methods include a credit card transaction or the vicarious execution settlement of accounts by a network operator, for example.
[0096]Although it was made to perform retrieval processing above via the Internet 2, this invention may be made to perform retrieval processing via the wire communication not only using this but cable TV etc., or the radio using a terrestrial wave or satellite broadcasting waves. In the case of radio, a portable telephone, PDA (Personal Digital Assistant), etc. may be sufficient as the terminal unit 3.
[0097]As mentioned above, the server system 1 can extract a search condition by repeating some questions to the ambiguous information demanded by the user. Therefore, an effect as taken below is acquired by using this invention.
(1) In an electronic video distribution system or an electronic music distribution system, an image or music can be searched from the vague impression which cannot necessarily carry out [keyword]-izing.
(2) The user can try listening it or preview an image or music to purchase before purchase.
(3) A user becomes possible [choosing and purchasing desired goods by interactive check] from a vague image.
[0098]Although a series of processings mentioned above can also be performed by hardware, they can also be performed by software. The computer by which the program which constitutes the software is included in hardware for exclusive use when performing a series of processings by software, Or it is installed in the personal computer etc. which can perform various kinds of functions, for example, are general-purpose, etc. from a recording medium by installing various kinds of programs.
[0099]. As shown in drawing 3, this recording medium is distributed apart from a computer in order to provide a user with a program. The magnetic disk 41 (a floppy disk is included) with which the program is recorded, the optical disc 42 (CD-ROM (Compact Disk-Read Only Memory).) . DVD (Digital Versatile Disk) is included. It is not only constituted by the package media which consist of the magneto-optical disc 43 (MD (Mini-Disk) is included) or the semiconductor memory 44, but, It comprises ROM22 with which a user is provided in the state where it was beforehand included in the computer and on which the program is recorded, the hard disk drive 29, etc.
[0100]In this specification, even if the processing serially performed in accordance with an order that the step which describes the program recorded on a recording medium was indicated is not

of course necessarily processed serially, it also includes a parallel target or the processing performed individually.
[0101]In this specification, a system expresses the whole device constituted by two or more devices.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the example of composition of the search system which applied this invention.
[Drawing 2]It is a block diagram showing the example of composition of the server system of drawing 1.
[Drawing 3]It is a block diagram showing the example of composition of the terminal unit of drawing 1.
[Drawing 4]It is a block diagram showing the example of composition of the retrieval server of drawing 2.
[Drawing 5]It is a figure explaining a retrieval query.
[Drawing 6]It is a figure showing a candidate list.
[Drawing 7]It is a flow chart explaining the message distribution processing of contents.
[Drawing 8]It is a flow chart following drawing 7.
[Drawing 9]It is a figure showing the display example of an initial entry screen.
[Drawing 10]It is a figure showing the display example of the screen which carries out additional questions.
[Drawing 11]It is a figure showing the display example of the screen which notifies search failure.
[Drawing 12]It is a figure showing the display example of the screen of a candidate list.
[Drawing 13]It is a figure showing the display example of the screen under an audition or preview.
[Drawing 14]It is a figure showing the display example of the User Information input screen.
[Drawing 15]It is a figure showing the display example of the screen which notifies disapproval.

[Drawing 16]It is a flow chart explaining retrieval processing.
[Drawing 17]It is a flow chart explaining accounting.
[Description of Notations]
1 A server system and 2 [An image / music server, and 14 / A fee collection server, 56 retrieving processors, and 57 retrieving databases] The Internet, 3-1, or 3-n A terminal unit and 11 A front-end processor, 12 retrieval servers, and 13

[Translation done.]

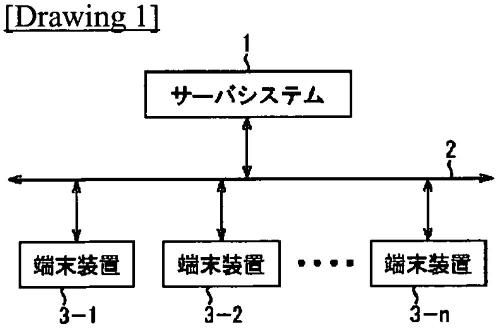
* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

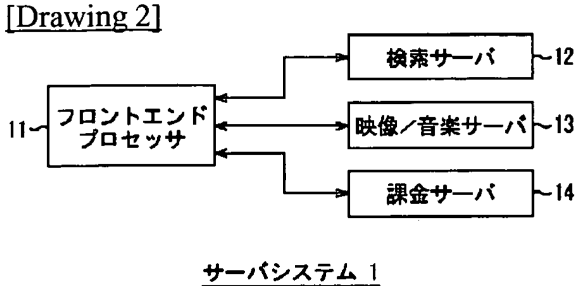
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

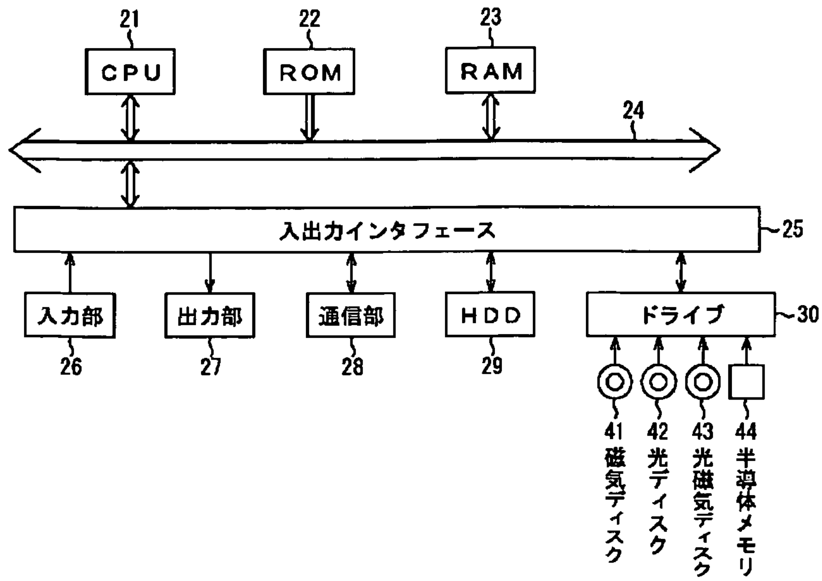
[Drawing 1]



[Drawing 2]



[Drawing 3]

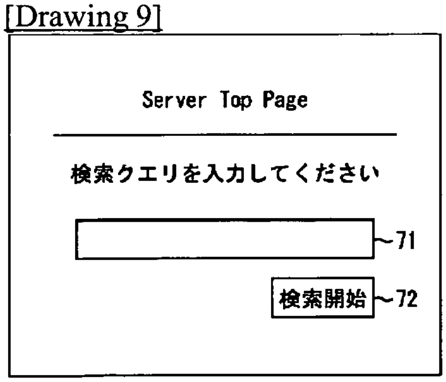


端末装置 3

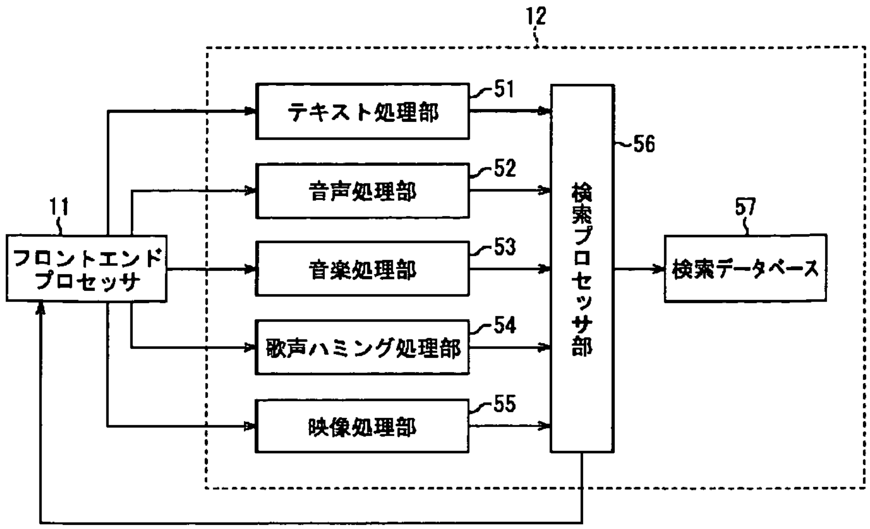
[Drawing 6]

	類似度	題名	クエリ位置
1	97%	ムーンリバー	3' 24"
2	88%	シェルブールの雨傘	1' 20"
3	83%	雨に唄えば	2' 30"
4	77%	虹の彼方に	0' 05"

[Drawing 9]



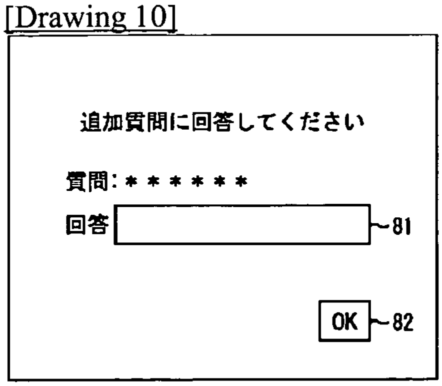
[Drawing 4]



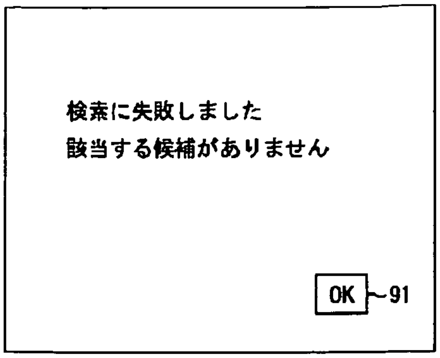
[Drawing 5]

	内容	形式
1.	楽曲のタイトル名、またはその一部分	テキスト·音声
2.	楽曲の演奏者名、またはその一部分	テキスト·音声
3.	楽曲の演奏者の性別、出身国名	テキスト·音声
4.	楽曲の作曲者名、またはその一部分	テキスト·音声
5.	楽曲の作詞者名、またはその一部分	テキスト·音声
6.	楽曲の指揮者名、またはその一部分	テキスト·音声
7.	楽曲のジャンル	テキスト·音声·音楽
8.	楽曲の歌詞または歌詞の一部分	テキスト·音声·歌声
9.	録音された楽曲、または楽曲の一部分	音楽
10.	ハミングまたは歌声による演奏、または一部分	歌声、ハミング
11.	その他楽曲に関連する情報(作曲年度、発売年度など)	テキスト·音声
12.	映像のタイトル名、またはその一部分	テキスト·音声
13.	映像の製作者名、またはその一部分	テキスト·音声
14.	映像に含まれる台詞、または台詞の一部分	テキスト·音声
15.	映像に含まれる主たる俳優の俳優名、またはその一部分	テキスト·音声
16.	録画された映像、または映像の一部分	映像·画像
17.	映像またはその一部を模擬または再現した映像または画像	映像·画像
18.	その他映像に関連する情報(製作年度、公開年度など)	テキスト·音声

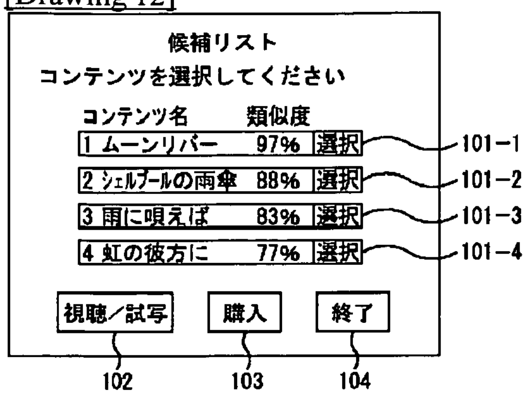
[Drawing 10]



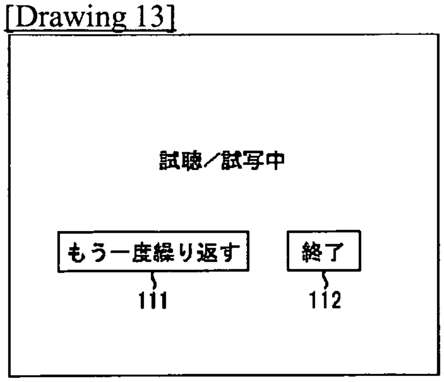
[Drawing 11]



[Drawing 12]

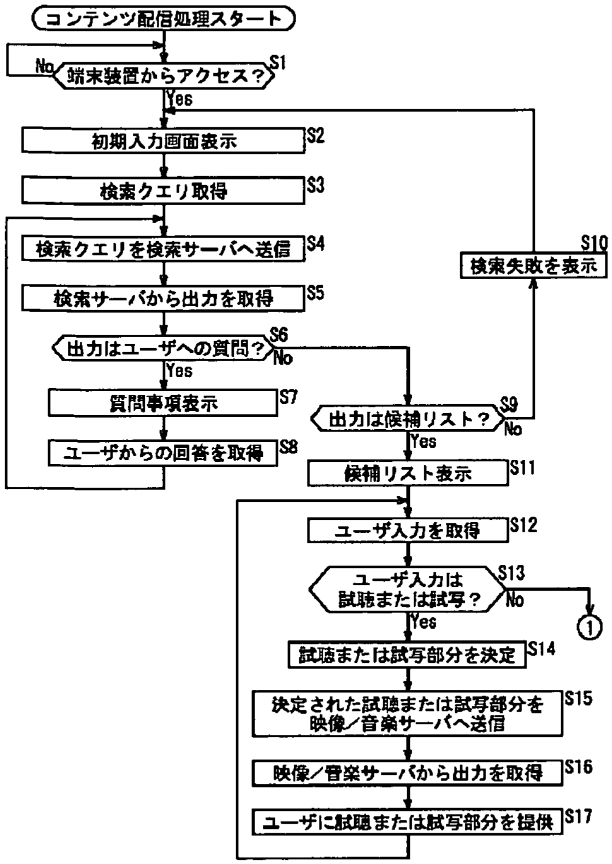


[Drawing 13]



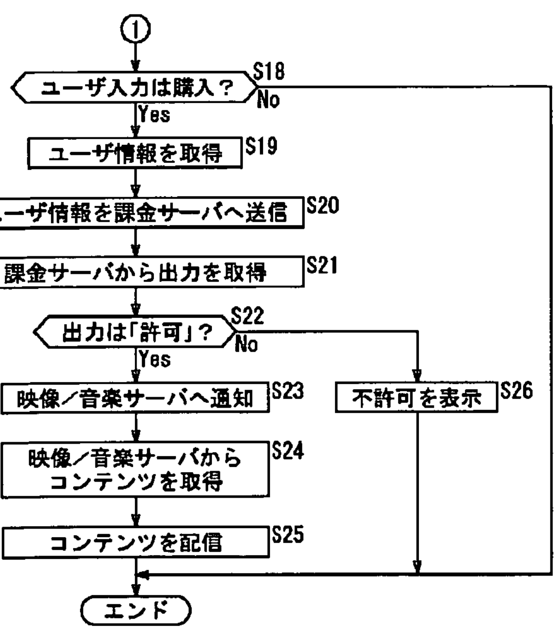
[Drawing 7]

(7-1)

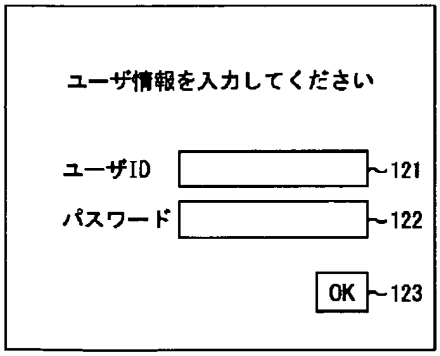


[Drawing 8]

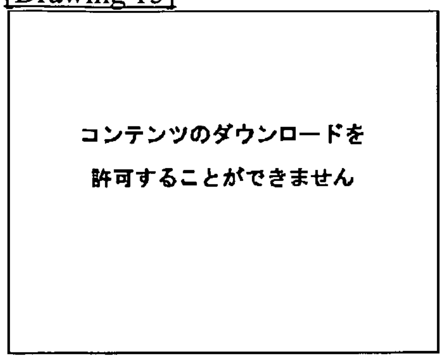
(7-2)



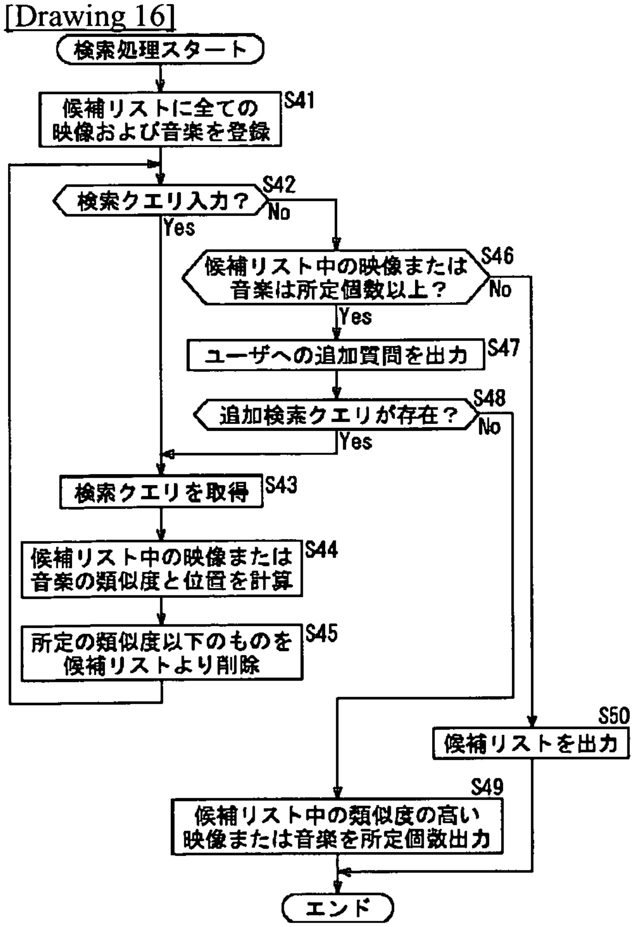
[Drawing 14]



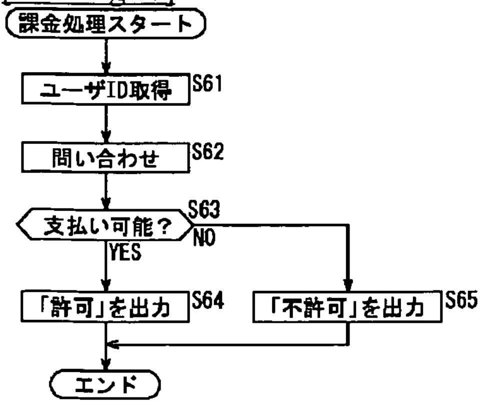
[Drawing 15]



[Drawing 16]



[Drawing 17]



[Translation done.]

PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2003-015684**
(43)Date of publication of application : **17.01.2003**

(51)Int.Cl. G10L 15/10
G10L 11/00
G10L 15/02
G10L 15/06
G10L 15/14

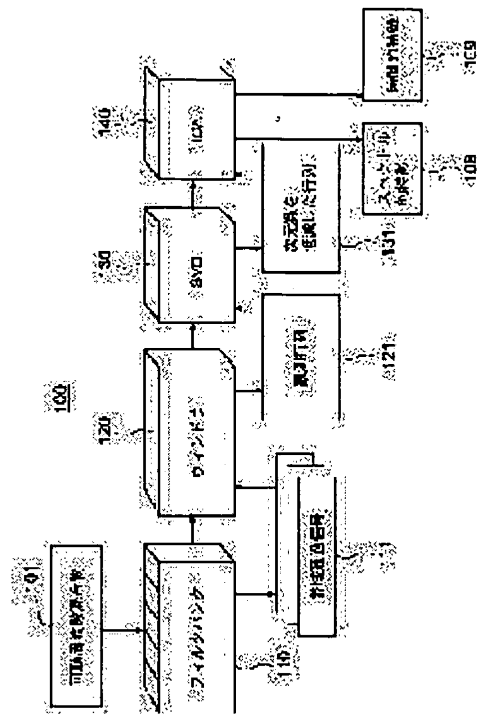
(21)Application number : **2002-146685**
(22)Date of filing : **21.05.2002**

(71)Applicant : **MITSUBISHI ELECTRIC RESEARCH LABORATORIES INC**
(72)Inventor : **MICHAEL A KASEI**

(30)Priority
Priority number : **2001 861808** Priority date : **21.05.2001** Priority country : **US**

(54) **METHOD FOR EXTRACTING FEATURE FROM ACOUSTIC SIGNAL GENERATED FROM ONE SOUND SOURCE AND METHOD FOR EXTRACTING FEATURE FROM ACOUSTIC SIGNAL GENERATED FROM A PLURALITY OF SOUND SOURCES**

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a computerized method for extracting features from acoustic signals generated from one or a plurality of sound sources.
SOLUTION: The acoustic signal are first windowed and filtered to produce a spectral envelope for each source. The dimensionality of the spectral envelope is then reduced to produce a set of features for the acoustic signal. The features in the set are clustered to produce a group of features for each of the sources. The features in each group include spectral features and corresponding temporal features characterizing each source. Each group of features is a quantitative descriptor that is also associated with a qualitative descriptor. Hidden Markov models are trained with sets of known features and stored in a database. The database can then be indexed by sets of unknown features to select or recognize like acoustic signals.



CLAIMS

[Claim(s)]
[Claim 1]A method characterized by comprising the following for extracting the feature from an audible signal generated from one sound source.
A step which carries out windowing of said audible signal, and filters it in order to generate a spectral envelope.
A step including the spectrum feature that are a step which reduces a number of dimension of said spectral envelope, and said one sound source is characterized by said group in order to generate 1 set of features, and the corresponding time feature.

[Claim 2]A method for extracting the feature from an audible signal generated from the one

sound source according to claim 1 which contains further a step which uses an outer product and carries out the multiplication of said spectrum feature and said time feature, since a spectrogram of said audible signal is reconstructed.
[Claim 3]A method for extracting the feature from an audible signal generated from the one sound source according to claim 1 which contains further a step which applies independent component analysis to said 1 set of features, in order to separate said feature in said group.
[Claim 4]A method for extracting the feature from an audible signal generated from the one sound source according to claim 1 which contains further a step which makes said spectral envelope logarithmic scale, is normalized by L2, and is made into a decibel graduation and the unit L2 norm, before reducing said number of dimension of said spectral envelope.
[Claim 5]A method characterized by comprising the following for extracting the feature from an audible signal generated from two or more sound sources.
A step which carries out windowing of said audible signal, and filters it in order to generate a spectral envelope.
A step which reduces a number of dimension of said spectral envelope in order to generate 1 set of features.
a group for each sound source of two or more of said sound sources -- a step including the spectrum feature that are a step which carries out clustering of said feature in said group, and said each sound source is characterized by said feature in said each group in order to generate the feature, and the corresponding time feature.

[Claim 6]A method for extracting the feature from an audible signal which the feature of each of said group is a quantitative description child of each of said sound source, and is generated from two or more sound sources according to claim 5 which contain further a step which associates a qualitative descriptor and said each quantitative description child in order to generate a category for said each sound source.
[Claim 7]A step of which a category in a database is composed as a sound source according to which a certain classification was classified, A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 6 which contain further a step which associates said each category in said database, and other at least one category with a related type link.
[Claim 8]A method for said category to extract the feature from an audible signal generated from two or more sound sources according to claim 7 stored in said database using a description definition language (DDL).
[Claim 9]A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 8 which define a base projection matrix to which a specific category in formation of DDL illustration reduces a series of specific logarithmic frequency spectrum of a sound source to a smaller number of dimension.
[Claim 10]A method for said category to extract the feature from an audible signal generated from two or more sound sources according to claim 6 containing cries and music other than an environmental sound, a background noise, sound effects, overlapping sounds, an animal sound, a sound, and a sound.
[Claim 11]A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 7 which contain further a step which combines an in general similar category in said database as a hierarchy of a class.
[Claim 12]A method characterized by comprising the following for extracting the feature from

an audible signal generated from two or more sound sources according to claim 6.
A specific quantitative description child is a harmonic envelope descriptor further.
A fundamental frequency descriptor.

[Claim 13]A step which divides into a state of a finite number said audible signal which said time feature describes a locus of said spectrum feature by the passage of time, and is generated by specific sound source based on said spectrum feature of corresponding, Since probability of transition in the following state is modeled when giving the present state, a step with which said each state is expressed according to continuous probability distribution, and, A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 5 which contain further a step with which said time feature is expressed by a transition matrix.
[Claim 14]Said continuous probability distribution is a 1xn vector of the average value m, and Gaussian distribution parameterized by the nxn covariance matrix K, however n is the number of the spectrum features in each spectral envelope, and it is the specific probability of spectral envelope x, [Equation 1]

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right]$$

A method for extracting the feature from the audible signal generated from two or more sound sources according to claim 13 given as be alike.
[Claim 15]A method characterized by comprising the following for extracting the feature from an audible signal generated from two or more sound sources according to claim 5 which contain further a step which stores said trained each Hidden Markov Model in a database.
A step which said each sound source is known, and uses a group of said feature in the case of said sound source of each known, and trains Hidden Markov Model to it.
A group of the related spectrum feature.

[Claim 16]A method characterized by comprising the following for extracting the feature from an audible signal generated from two or more sound sources according to claim 5 which contain further a step which stores trained each Hidden Markov Model.
A step which 1 set of audible signals belong to a known category, and extracts a spectrum base for said audible signal.
A step which trains Hidden Markov Model using said time feature of said audible signal.
Said related spectrum basis function.

[Claim 17]A step characterized by comprising the following, and since said strange sound source is specified, A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 15 which contain further a step which chooses one of stored Hidden Markov Model which suits a group of said strange feature best.
A step which generates a strange audible signal from a strange sound source.
A step which carries out windowing of said strange signal, and filters it in order to generate a strange spectral envelope.
The strange spectrum feature that are a step which reduces a number of dimension of said strange spectral envelope, and said strange sound source is characterized by said group in order to generate 1 set of strange features.
The corresponding strange time feature.

[Claim 18]A method for extracting the feature from an audible signal generated from two or more sound sources according to claim 17 chosen since said two or more stored Hidden Markov Model specifies two or more strange sound sources in general similar to said strange sound source.

[Translation done.]

* NOTICES *

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the method of recognizing and carrying out indexation of the audible signal, and searching for it, in detail about the field of acoustic signal processing generally.
[0002]
[Description of the Prior Art]About extracting the feature of an environmental sound and an ambient sound until now, research was hardly made. The audible signal style of a great portion of conventional technology has been concentrated on human being's sound and music. There is no style for which it was suitable for many sound effects it is heard on the other hand in a movie, television, a video game, and virtual environment, such as umbrella umbrella ******** and a sound into which water flows, that a footstep, a traffic sound, the sound that shuts a door with BATAN, the sound of a laser gun, the sound struck hard, the sound to throw, a thunder, and a leaf are. Generally it is dramatically difficult for these environmental audible signals to extract the feature compared with a sound and music. Because, when those signals are many, it is because the ingredient of piled-up a large number and a high order structural ingredient like repetition and dispersion are included including noise.
[0003]One specific application gestalt which can use such an expression method is graphic processing. The method for extracting, compressing, looking for and classifying an image object is available. For example, please refer to various MPEG standards. Except when the subject of an "audio frequency belt sound" is a sound, the method of processing the subject of such an "audio frequency belt sound" does not exist. For example, while John Wayne shoots 6 running fire handguns, in order to identify the position of all the images along which he is running at full speed on the horse, to search an image library may be desired. To be sure, it is possible to specify John Wayne or a horse visually. However, it is dramatically difficult to sort out the sound of

rhythmical PAKAPPAKATSU of a horse which runs at full speed, and the intermittent percussion sound of a revolver. By recognizing the phenomenon of an audio frequency belt sound, the operation within an image can be described in detail.

[0004]Another application gestalt which can use the style is composition of a sound. In order to compound a sound and to be able to generate by methods other than trial and error, the feature of a sound must be specified before that.

[0005]In conventional technology, the expression for sounds other than a sound reproduces the tone of a specific musical instrument general, for example, It has concentrated on sounds other than a sound of a specific class like identifying the specific musical instrument, distinguishing the sound of a submarine from the sound of the surrounding sea, and recognizing the underwater mammals by the cry. These application gestalten need the acoustical feature of the specific arrangement which is not generalized more than a specific application gestalt, respectively.

[0006]In addition to these specific application gestalten, other researches have been concentrated on developing analysis expression of the acoustical sight generalized. This research came to be known as "auditory scene analysis by calculation." These systems originate in the complexity of the algorithm, and need great computation work. Typically, a method [heuristics / from artificial intelligence and various reasoning methods] is used for those systems.

[0007]Although such a system gives the useful discernment to the arduous problem about sound expression, it is not once shown about a classification and composition of the audible signal in the state where the performance of the system was mixed that it is satisfactory.

[0008]In another application gestalt, indexation of the medium of an audio frequency belt sound including the phenomenon of the sound of the broad range containing cries and music other than an environmental sound, a background noise, sound effects (sound effect), an animal sound, a sound, and a sound can be carried out using expression of a sound. The sound recognition tool for looking for the medium of an audio frequency belt sound can be designed now using the index extracted automatically by this. Semantic description of the contents or the similarity to target reference of an audio frequency belt sound can search for a sound track including the contents of many like a movie or a report program using these tools. For example, to pinpoint the position of all the image clips with which a lion barks or an image raises a cry is desired.

[0009]There is approach in which much realization to automatic classification and indexation is possible. Wold (IEEE Multimedia, pp.27- 36, 1996) etc., "Martin etc. Musical instrument. identificationa pattern-recognition approach: " (Presented at the 136 th Meeting of the Acoustic Society.) of America, Norfolk, VA, and 1998 indicate the strict classification for a musical instrument. "Zhang etc. Content-based. classification and retrieval. of. audio." (SPIE 43rd Annual Meeting, Conference on Advanced Signal Processing Algorithms, Architectures and ImplementationsVIII,.) 1998 indicates the system which trains a model using spectrogram data, "Boreczky etc. A hidden. Markov model framework. for video segmentation using audio and image features" (Proceedings of ICASSP'98, pp.3741- 3744, 1998) uses a Markov model.

[0010]

[Problem(s) to be Solved by the Invention]Carrying out indexation of the audio frequency belt sound medium, and looking for it relates to MPEG-7 standards which newly because of multimedia appeared especially closely. The standards need the interface integrated to the class of a general sound. The compatibility of an encoder is one element about a design. In that case, the database of the "sound" which has an index provided by one embodiment can be compared with the database extracted by different embodiment.

[0011]

[Means for Solving the Problem]By a computerized method, the feature is extracted from an audible signal generated from one or more sound sources. Windowing of the audible signal is carried out first, it is filtered, and a spectral envelope to each sound source is generated. Then, a number of dimension of a spectral envelope is reduced and the feature which is 1 set for the audible signal is generated. a group [as opposed to / clustering of the feature in the group is carried out, and / each sound source] -- the feature is generated. The feature in each group includes the spectrum feature by which each sound source is characterized, and the corresponding time feature.
[0012]The feature of each group is a quantitative description child, and a quantitative description child is related also with a qualitative descriptor. Hidden Markov Model is trained by group of the known feature, and is stored in a database. Indexation can be carried out by group of the strange feature in order that the database may choose or recognize a similar audible signal in that case.
[0013]
[Embodiment of the Invention]Drawing 1 shows the method 100 for extracting the spectrum feature 108 and the time feature 109 from the mixture 101 of a signal by this invention. The method 100 of this invention is the purpose of classifying a sound source, and it can be used in order to clarify the feature from the recorded sound and to extract it, and in order to change and reuse the purpose in the application gestalt of structurized multimedia like composition of a parameter (re-purpose). The method can be used also in order to extract the feature from the mixture of many dimensions further, the mixture of other linearity, and. The mixture is obtained from one sound source or the sound source of a large number like stereo sound sources.
[0014]In order to extract the feature from the recorded signal, the method by this invention uses the statistical technique based on independent component analysis (ICA). ICA conversion generates rotation of the base of the time-frequency observation matrix 121 using the contrast function defined by the cumulative extension up to the 4th order.
[0015]The base ingredient generated as a result is statistically [as possible] independent, and clarifies the each feature within the mixture sound source 101, for example, the feature of the structure of a sound. The new signal which classifies a signal or has the feature which can be predicted can be specified using such characteristic structures.
[0016]The expression by this invention can compound behavior of many sounds from the feature of a small group. The expression by this invention can compound the characteristic of acoustical subjects, such as complicated acoustical phenomenon structures of colliding, such as bouncing, striking, and grinding, and material, a size, and shape.
[0017]In the method 100, the audio frequency belt sound mixture 101 is first processed by the bank 110 of a logarithmic filter. Each filter generates the band pass signal 111 for a predetermined frequency range. Typically, the band pass signal 111 of 40-50 is generated, and in a low-pass frequency range, many signals are generated from the frequency range of a high region so that the frequency response characteristic of human being's ear may be imitated. In an exception method, as the filter, constant Q (CQ) or a wavelet filter bank can be used, or the filter can be arranged like [in the short-time Fast Fourier Transform expression (STFT)] at linearity.
[0018]In Step 120, for example, each band pass signal is short, "windowing" is carried out to 20msec segment, and an observation matrix is generated. Each procession can also contain hundreds of samples. The details of Steps 110 and 120 are shown in drawing 2 and drawing 3 still in detail. Please care about that windowing can be performed before filtering.
[0019]In Step 130, singular value decomposition (SVD) is applied to the observation matrix 121,

and the procession 131 which had the number of dimension reduced is generated. SVD was indicated for the first time by geometrician belt RAMI of Italy in 1873. Singular value decomposition is clear generalization of principal component analysis (PCA). mxn singular-values-of-a-matrix decomposition is arbitrary factorization of the following forms.
[0020]X=UsigmaV ^{T}[0021]However, U has an orthogonal matrix of mxm, U has a regular rectangular cross sequence, V is an orthogonal matrix of nxn, and sigma is a diagonal matrix of mxn of the singular value in which i has ingredient $sigma_{ij}$=0 when not equal to j.
[0022]As one advantage, in contrast with PCA, SVD can decompose a non-square matrix and, thereby, an observation matrix can be directly decomposed in a spectrum or the direction of one of time, without needing to calculate a covariance matrix. Since SVD decomposes a non-square matrix directly, without needing to ask for a covariance matrix, the base generated as a result cannot be easily influenced by PCA to the problem of a dynamic range.
[0023]The method of this invention applies the independent component analysis (ICA) of an option to the procession 131 which had the number of dimension reduced in Step 140. ICA using the repetitive on-line algorithm based on the false neuro-architecture for blind signal separation is known well. Much neural network architecture for solving an ICA problem these days is proposed. "For example, it was given to Sejnowski on January 17, 1995 Adaptive system for broadband multisignal discrimination in a channel. Please refer to U.S. Pat. No. 5,383,164 of a title called with reverberation."
[0024]ICA generates the spectrum feature 108 and the time feature 109. The spectrum feature expressed as a vector corresponds to the point estimate of the ingredient which is independently statistically in a segmentation window. Similarly the time feature is expressed as a vector and describes deployment of the spectral component in the process of the segment.
[0025]It is combined using the outer product of a vector and each set of a spectrum and a time vector can reconstruct the partial spectrum for a given input spectrum. When reversible, these spectra can presume the independent segment-of-time signal, so that filter bank expression may be so. In the case of the each independent ingredient described in the method, the procession of the compatibility score for the ingredient in a former segment is available. An ingredient can be pursued now over time by presuming the continuous high correspondence of possibility thereby most. Only while seeing the front in time, it is equal to a backward interchangeability procession.
[0026]Each signal component in an audio frequency belt sound track can be presumed using independent component disassembly of the track of an audio frequency belt sound. When the signal procession (mixture of N linearity of N sound sources) of the number of the whole floor cannot be used, it is hard to deal with a separation problem, but the approximation to the sound source in the bottom can be given by using the independent component of the short time section of frequency domain expression. These approximation can be used for a classification, recognition operation, and comparison between sounds.
[0027]As shown in drawing 3, temporal modulation distribution (TFD) can be normalized with the power spectral density (PSD) 115, in order to make small contribution of a lower frequency component which conveys more energies in some sound fields.
[0028]Drawing 4 and drawing 5 show the time and spatial decomposition about the percussion instrument played in a regular rhythm, respectively. By the structure which can be observed, the articulate ingredient of the broadband corresponding to a shake and the horizontal layer system corresponding to the singing of a metal shell become clear.
[0029]Application gestalt this invention for the acoustical feature of a sound can be used in many application gestalten. It can be considered that the extracted feature is a disengageable ingredient

showing a peculiar structure in a sound-source mixture of an acoustical mixture. Since the ingredient is recognized or specified, the extracted feature can be compared with 1 set of transcendental classes determined by pattern recognition art. These sorters can be in the field of the analytical model by a phoneme, sound effects, a musical instrument, the animal sound, or other arbitrary corpora. The extracted feature can be individually re-compounded using a reverse filter bank, and, thereby, can attain "purification" of the acoustical mixture of a sound source. The use of an example is separating a singer, a drum, and a guitar from the recorded sound, in order to change the purpose and to reuse some ingredients, or in order to analyze a musical structure automatically. Another example is separating an actor's voice from a background noise and passing a clear audio signal to a speech recognizer, in order to make title translation of the movie automatically.

[0030]The spectrum feature and the time feature can be individually taken into consideration, in order to identify the various characteristics of the acoustical structure of the subject of each sound in a mixture. The time feature can explain behavior of bouncing, breaking, striking, etc. to the ability of the spectrum feature to explain material, a size, and the characteristic like shape. In this way, it is distinguishable from that a glass bounces or striking earthenware to strike a glass. The extracted feature can be changed and re-compounded in order to generate the synthetic example with which the sound of the sound source was changed. When an input sound is a phenomenon of one sound including two or more acoustical features, such as striking a glass, each feature can be controlled for re-composition. This is useful because of the application gestalt of the medium by models, such as generating a sound in virtual environment.

[0031]Using indexation, search, and this invention, indexation of the big multimedia database including the sound of the type with which many differ, for example, sound effects, an animal sound, a musical instrument, a sound, the overlapping sounds, an environmental sound, a masculine sound, and a feminine sound can be carried out, and it can also search for it.

[0032]Generally in this context, description of a sound is divided into the qualitative description using the character by two types, i.e., a category label, and the quantitive description using a stochastic model state. A category label provides the qualitative information about the contents of the sound. The description in this form is suitable for the application gestalt of reference in a character like the arbitrary processing tools which use the Internet search engine or an alphabetic field.

[0033]By contrast, the quantitative description child can use including the compact information about the segment of an audio frequency belt sound for numerical evaluation of the similarity of a sound. For example, in video or audio sound recording, a specific musical instrument is discriminable using these descriptors. Qualitative and a quantitative description child suits the application gestalt of illustration reference search of an audio frequency belt sound.

[0034]While segmenting in a class a sound recognition descriptor and the audio frequency belt sound by which recording mode qualitative descriptor sound recording was carried out, to acquire the semantic information related about the contents is desired. For example, by recognizing the scream in an image sound track, fear or danger can be directed and a comedy can be directed by a laughing voice. The sound can direct existence of people and, so, the video segments to which these sounds belong can be used as a candidate at the time of looking for a clip including people. The category and system-of-classification descriptor of a sound provide the means for composing a category concept to the layered structure which makes possible this type of complicated related type searching method.

[0035]the category of a sound -- as shown in drawing 6 for the easy classification 600, since the

category of a sound is named, the recording mode (DS) is used. the sound at which a dog barks as an example -- as a subcategory -- "-- it barks and has voice"611 -- qualitative -- category label "dog" 610 can be given. Furthermore, "groan" 612 or "******" 613 can be made into the desirable subcategory of a "dog." Although the first two subcategories are associated closely, the 3rd subcategory is a phenomenon of a completely different sound. So, drawing 6 shows that four categories are composed by the classification which has "dog" 610 as a root node. Each category has at least one related link 601 to another category in the classification. By initial setting, it is considered that the category accommodated is the category (NC) 601 narrower than the accommodated category. however -- this example -- a "groan" -- "-- although it barks and is synonymous with voice" in general, it defines as a thing not more desirable than it. In order to acquire such a structure, the following relations are defined as a part of recording mode of this invention.
[0036]A category larger than BC- means that the category associated is more common in a meaning than in the accommodated category. A category narrower than NC- means that the category associated is more restrictive in a meaning than in the accommodated category. US - Since it is more desirable than the present category, the present category and the category with which homonymy is related in general are used. UF - Use of the present category is more preferred than the category with which homonymy is related mostly. RC - The category associated is related with homonymy and the category accommodated although it is not homonymy and a larger or narrower category to some extent.
[0037]The following XML-schema codes show how the qualitative recording mode for the categorization method shown in drawing 6 is illustration-ized using a description definition language (DDL).
[0038]
[Equation 2]

```
<SoundCategory term="1" scheme="DOGS">
  <Label>Dogs</Label>
  <TermRelation term="1.1" scheme="DOGS">
    <Label>Bark</Label>
    <TermRelation term="1.2" scheme="DOGS" type="US">
      <Label>Woof</Label>
    </TermRelation>
  </TermRelation>
  <TermRelation term="1.3" scheme="DOGS">
    <Label>Howl</Label>
  </TermRelation>
</SoundCategory>
```

[0039]Both a category and a method attribute provide the peculiar identifier which can be used in order to refer to the following categories and classifications from a quantitative description method like a stochastic model that are indicated still in detail. A label descriptor gives the significant clever semantic label for each category, and relation descriptors describe the relation in the category of the classification by this invention.
[0040]As shown in system-of-classification drawing 7, a category can be combined with the

system of classification 700 with a related link, and more abundant classifications can be created. for example, -- "-- it barks, voice"611 are a subcategory of "dog" 610, and "dog" 610 are a subcategory of "pet" 701. As for it, category "cat" 710 are the same. The cat 710 has category "cry" 711 of a sound, and "sound with which throat is sounded" 712. The following is an example of an easy system of classification for a "pet" containing two categories "dog" and a "cat."

[0041]In order to carry out this system of classification by extending a method defined beforehand, the 2nd method to which a name of a "cat" was given is illustration-ized as follows.

[0042]

[Equation 3]

```
<SoundCategory term="2" scheme="CATS">
  <Label>Cats</Label>
  <TermRelation term="2.1 scheme="CATS">
    <Label>Meow</Label>
  </TermRelation>
  <TermRelation term="2.2" scheme="CATS">
    <Label>Purr</Label>
  </TermRelation>
</SoundCategory>
```

[0043]In order to combine these categories here, the system of classification called a "pet" is illustration-ized with reference to the method defined beforehand.

[0044]

[Equation 4]

```
<ClassificationScheme term="0" scheme="PETS">
  <Label>Pets</Label>
  <ClassificationSchemeRef scheme="DOGS"/>
  <ClassificationSchemeRef scheme="CATS"/>
</ClassificationScheme>
```

[0045]Here, the system of classification called a "pet" contains the category "pet" additional as a route including all the category elements of a "dog" and a "cat." In the case of the application gestalt of character indexation, the above qualitative classifications are enough.

[0046]The following sections indicate the quantitative description child for a classification and indexation who is used with a qualitative descriptor and can form the indexation and searching engine of a perfect sound.

[0047]A quantitative description consonant recognition quantitative description child describes the feature of the audible signal which will be used with a statistical sorter. The sound recognition quantitative description child can use for recognition of the general sound containing sound effects and a musical instrument. the descriptor suggested -- in addition, other arbitrary descriptors defined in the structure of an audio frequency belt sound can be used for a classification.

[0048]The feature most extensively used for the classification of the audio frequency band-spectrum base feature sound is a power-spectrum slice or expression by a spectrum like a frame.

Typically, each spectrum slice is a vector of n dimension, and n is the number of spectrum channels and has a channel of the data up to 1024 channels. With logarithmic frequency spectrum which is expressed by the structural description child of an audio frequency belt sound, a number of dimension can be reduced to about 32 channels. So, generally, the feature drawn by the spectrum originates in a high number of dimension, and is incompatible with a stochastic model sorter. The probability sorter operates the best with a number of dimension smaller than ten dimensions.

[0049]So, a basis function of the number of low dimensions generated by the above and the following singular value decomposition (SVD) is preferred. In that case, an audio frequency belt sound spectrum base descriptor is a container for a basis function used in order to project the spectrum on subspace of a low dimension for which it was suitable for a stochastic model sorter.

[0050]This invention determines a base for each class of a sound, and a subclass. The base acquires the statistical most regular feature of a feature space of a sound. Reduction of a number of dimension is performed by projecting a spectrum vector as mentioned above to a procession of a basis function drawn from data. The number of lines corresponds to the length of a spectrum vector, and a basis function is stored in a sequence of a procession corresponding to the number of basis functions in the number of sequences. Base projection is a procession product of a spectrum and a base vector.

[0051]Spectrogram drawing 8 reconstructed from a basis function shows the spectrogram 800 reconstructed from four basis functions by this invention. The concrete spectrogram is a thing for "pop" music. The left-hand side spectrum base vector 801 is combined with the base projective vector 802 using an outer product of a vector. A procession of an outer product generated as a result, respectively is added, and a final reconstruction thing is generated. A basis function is chosen so that information may be made into the maximum in a small number of dimension from the original data. For example, the basis function can respond to a statistically independent ingredient which corresponds to the feature of not correlating [which is extracted using principal component analysis (PCA) or Karhunen-Loeve conversion (KLT)], or is extracted by independent component analysis (ICA). When a secondary statistic, i.e., covariance, understands KLT or the Hotelling conversion, it is desirable negative correlation conversion. This reconstruction is indicated still in detail with reference to drawing 13.

[0052]In order to achieve the purpose of a classification, a base for all the classes is drawn. In this way, classifying space contains a statistically remarkable ingredient of the class. The following DDL illustration-ization defines a base projection matrix which reduces a series of logarithmic frequency spectrum of 31 channels to five dimensions.

[0053]

[Equation 5]

```xml
<AudioSpectrumBasis loEdge="62.5" hiEdge="8000" resolution="1/4 octave">
  <Basis>
    <Matrix dim="31 5">
    0.26 -0.05 0.01 -0.70 0.44
    0.34 0.09 0.21 -0.42 -0.05
    0.33 0.15 0.24 -0.05 -0.39
    0.33 0.15 0.24 -0.05 -0.39
    0.27 0.13 0.16 0.24 -0.04
    0.27 0.13 0.16 0.24 -0.04
    0.23 0.13 0.09 0.27 0.24
    0.20 0.13 0.04 0.22 0.40
    0.17 0.11 0.01 0.14 0.37
   ...
      </Matrix>
  </Basis>
</AudioSpectrumBasis>
```

[0054]Low edge, high edge, and a resolution attribute give the interval of the spectrum channel in the bottom frequency limit, the upper part frequency limit, and octave-band notation of a basis function. In the classification structure by this invention, the basis function for all the classes of a sound is stored with the stochastic model for the class.

[0055]The features used for the feature sound recognition of sound recognition can be collected, and the one recording mode which can be used for various different application gestalten can be adopted. In a classification of the sound acquired, many types, for example, sound-effects libraries, of the sound, and the sample disk of a musical instrument, the audio frequency belt sound spectrum projection descriptor of initial setting plays a role good.

[0056]The base feature is drawn from the above audio frequency belt sound spectral envelope extraction processes. An audio frequency belt sound spectrum projection descriptor is a container for the feature which reduced a number of dimension similarly obtained by the projection of a spectral envelope to 1 set of basis functions as mentioned above. For example, an audio frequency belt sound spectral envelope is extracted by sliding window FFT analysis with re-sampling to a frequency band arranged by logarithm. In a desirable embodiment, analysis frame periods are 10msec. However, a slide extraction window of 30msec temporal duration is used in a Hamming window. 30msec interval provides sufficient spectral resolving power, and it is chosen so that a channel of the beginning of the 62.5-Hz width of an octave-band spectrum may be decomposed in general. A size of an FFT-analysis window is a sample number of involution of 2 big next. As for this, in 30msec, 960 samples exist at 32 kHz, but it means that FFT will be performed in 1024 samples. In 30msec, 1323 samples exist at 44.1 kHz, but FFT will be performed in 2048 samples and a sample outside a window is set as 0.

[0057]Drawing 9 a and drawing 9 b show the base projection 911-913 in the case of the frequency index 920 for the "laughing voice" spectrogram 1000 in the three spectrum base ingredients 901-903, drawing 10 a, and drawing 10 b of a case of the time index 910 generated. A form here is similar to form shown in drawing 4 and drawing 5. Drawing 10 a shows a spectrogram of logarithmic scale of a laughing voice, and drawing 10 b shows what

reconstructed a spectrogram. Which drawing plots time and a frequency index on a x axis and the y-axis, respectively.

[0058]In addition to a base descriptor, a sorter can be defined using the special characteristic of a class of a sound like a harmonic envelope for a musical instrument classification used in many cases, and the fundamental frequency feature using another quantitative description child's big sequence.

[0059]The one convenience of number-of-dimension reduction which is made by this invention is that arbitrary descriptors based on 1 set of descriptors in which scaling is possible can add to a spectrum descriptor with the same sampling rate. A suitable base is calculable to the whole group of the extended feature like a base based on a spectrum.

[0060]Another application gestalt for the feature recording mode of sound recognition by spectrogram abstract-ized this invention using a basis function is an efficient spectrogram expression. Visualization and in order to abstract-ize, audio frequency belt sound spectrum base projection and the audio frequency belt sound spectrum base feature can be used for a spectrogram as dramatically efficient memory mechanisms.

[0061]Since a spectrogram is reconstructed, this invention uses the formula 2 indicated in detail by the following. The formula 2 constitutes a two-dimensional spectrogram from a cross product of each basis function shown also in <u>drawing 8</u> as mentioned above, and its corresponding spectrogram basic projection.

[0062]Since the stochastic model recording mode finite state model spectrum feature is changed over time, a phenomenon of a sound is dynamic. This very big time change gives characteristic "fingerprint" for recognition to an audible signal. So, a model of this invention divides into a limited state number an audible signal generated by class of a specific sound source or a sound. The division is based on the spectrum feature. Each sound is described by locus passing through this state space of those sounds. This model is indicated in detail by the following about drawing 11 a and drawing 11 b. Each state can be expressed by continuous probability distribution like Gaussian distribution.

[0063]Dynamic behavior of a class of a sound passing through state space is expressed by transition matrix of kxk which describes probability of transition in the following state when giving the present state. The transition matrix T models probability of transition in the state j in the time t from the state i in time t-l. Early state distribution is kx1 vector of probability, and is typically used also in a finite state model. The k-th element of this vector is probability which is in the state k in the first observation frame.

[0064]Since a state is modeled during a classification of a sound, Gaussian distribution type multi-dimension Gaussian distribution is used. Gaussian distribution is parameterized by a 1xn vector of the average value m, and the covariance matrix K of nxn. However, n is the number of the features in each observation vector. When a gauss parameter is given, a formula for calculation of probability to the specific vector x is as follows.

[0065]

[Equation 6]

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{x}-\mathbf{m})^T \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m}) \right]$$

[0066]Continuation Hidden Markov Model is a finite state model which has a continuous-probability-distribution model for state observation probability. The following DDL illustration-ization is examples of use of the stochastic model recording mode for expressing the

continuation Hidden Markov Model which has a gauss state. Below as for the decimal point, in this example, the floating point number is rounded off by double figures only for the purpose of a display.
[0067]
[Equation 7]


```
<ProbabilityModel xsi:type="ContinuousMarkovModelType" numberStates="7">
<Initial dim="7">
0.04 0.34 0.12 0.04 0.34 0.12 0.00 </Initial>
<Transitions dim="7 7">
0.91 0.02 0.00 0.00 0.05 0.01 0.01
0.01 0.99 0.00 0.00 0.00 0.00 0.00
0.01 0.00 0.92 0.01 0.01 0.06 0.00
0.00 0.00 0.00 0.99 0.01 0.00 0.00
0.02 0.00 0.00 0.00 0.97 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.98 0.01
0.02 0.00 0.00 0.00 0.00 0.02 0.96
</Transitions>
<State><Label>1</Label></State>
<!—State 1 Observation Distribution -- >
<ObservationDistribution xsi:type="GaussianDistributionType">
<Mean dim="6">
5.11 -9.28 -0.69 -0.79 0.38 0.47
</Mean>
<Covariance dim="6 6">
1.40 -0.12 -1.53 -0.72 0.09 -1.26
-0.12 0.19 0.02 -0.21 0.23 0.17
-1.53 0.02 2.44 1.41 -0.30 1.69
-0.72 -0.21 1.41 2.27 -0.15 1.05
0.09 0.23 -0.30 -0.15 0.80 0.29
-1.26 0.17 1.69 1.05 0.29 2.24
</Covariance>
<State><Label>2</Label></State>
<!—Remaining states use same structures-- >
<\PobabilityModel>
```


[0068]In this example, a "stochastic model" is illustration-ized as a Gaussian distribution type drawn from a base stochastic model class.
[0069]By the method by this invention, the tool has so far [sound recognized model recording mode] been separated, without completely using the structure of an application gestalt. The following data types combine an above-mentioned descriptor and recording mode, and make them the structure where it was unified for the classification of a sound, and indexation. Indexation of the segment of a sound can be carried out with the category label based on the output of a sorter. A probability model parameter can be used for the indexation of the sound in a database. To carry out indexation by a model parameter like a state is needed by the illustration reference application gestalt, when a reference category is strange, or when a collation judgment standard narrower than the range of a category is needed.
[0070]The sound recognized model sound recognized model recording mode specifies the stochastic model of the class of a sound like Hidden Markov Model or a gauss mixing model. the following examples -- drawing 6 -- "-- barking -- the Hidden Markov Model of the voice" sound category 611 -- **** -- it is-izing. The stochastic model and the related basis function for the class of the sound are similarly defined as the case of the example indicated previously.
[0071]
[Equation 8]

```
<SoundRecognitionModel id="sfx1.1" SoundCategoryRef="Bark">
<ExtractionInformation term="Parameters" scheme="ExtractionParameters">
<Label>NumStates=7, NumBasisComponents=5</Label>
</ExtractionInformation>
<ProbabilityModel xsi:type="ContinuousMarkovModelType" numberStates="7">
        ... <!-- see previous example -->
</ProbabilityModel>
<SpectrumBasis loEdge="62.5" hiEdge="8000" resolution="1/4 octave">
        ... <!-- see previous example -->
</SpectrumBasis>
</SoundRecognitionModel>
```


[0072]sound-models state Pass -- this descriptor describes the dynamic state path of a sound through that model with reference to a finite state stochastic model. Indexation can be carried out [sound] in two modes segmenting a sound in the model state, or by sampling a state path at a regular interval. As for the temporal duration of the segment, in the case of the 1st, each audio frequency belt sound segment directs the effective temporal duration for the state including the reference to one state. In the case of the 2nd, a sound is described by the index of a sampled series which refers to a model state. The sound category which has comparatively long state temporal duration is efficiently described using one segment and 1 state approach. The sound which has comparatively short state temporal duration is described still more efficiently using a series of sampled state indices.
[0073]Drawing 11 a shows the logarithmic spectrogram (frequency versus time) 1100 of the dog **** vocal sound 611 of drawing 6. Drawing 11 b shows a sound-models state path sequence in the state where drawing 11 a barked and it let continuation Hidden Markov Model for a voice model pass covering the same time interval. In drawing 11 b, a x axis is a time index and the y-axis is a state index.
[0074]Sound recognition sorter drawing 12 shows a sound recognition sorter which uses the one database 1200 for all the required ingredients of a sorter. The sound recognition sorter describes a relation between many stochastic models, and, thereby, defines ontology of a sorter. for example, a class of an extensive sound like [so that a hierarchical recognizer may be indicated in the case of drawing 6 and drawing 7] an animal in a root node -- moreover -- Dog: barking in a leaf node -- voice and a cat:cry, and ** -- a finer class [like] can be classified. Using descriptor method structure of a graph, this method defines a correspondence relation between ontology of a sorter, and classification of a category of a sound, and in order that hierarchical sound models may extract category description in the case of given classification, it is used.
[0075]Drawing 13 shows the system 1300 for constituting a database of a model. A system shown in drawing 13 is an extended type of a system shown in drawing 1. Here, before filtering in order to extract a spectral envelope, windowing of the input acoustic signal is carried out. The system can incorporate the audio frequency belt sound inputting 1301 in a form of an audio file of WAV form, for example. The system extracts the audio frequency belt sound feature from a file, and trains Hidden Markov Model in these features. In the case of a class of each sound, a directory of a specimen of a sound is used for the system. Hierarchical directory structure defines ontology corresponding to desired classification. In the case of each directory of the ontology, one Hidden Markov Model is trained.
[0076]The system 1300 of audio frequency belt sound feature extraction drawing 13 shows a method for extracting an audio frequency belt sound spectrum basis function and the feature from an audible signal as mentioned above. The input acoustic signal 1301 is generable by one sound source, for example, a person, animal, a musical instrument, or many sound sources, for

example, people and an animal, many musical instruments or composite tone. In the case of the latter, an audible signal is a mixture. Windowing of the input acoustic signal is first carried out to 10msec frame (1310). In drawing 1, an input signal should care about that band pass filtering is carried out before windowing. Here, windowing of the audible signal is carried out first, it is filtered after that (1320), and extracts frequency spectrum (short-time logarithmic-in-frequency spectrum) logarithmic [short-time]. Filtering performs time-frequency power-spectrum analysis like short time Fourier transformation (squared-magnitude) which squared a size. The result is a procession which has M frames and the frequency (frequency bins) of N pieces. The spectrum vector x is a line of this procession.

[0077]Step 1330 performs normalization of logarithmic scale. Each spectrum vector x is changed into the decibel graduation 1331 from a power spectrum by $z=10\log_{10}(x)$. Step 1332 determines the L2 norm of a vector element as follows.

[0078]

[Equation 9]

$$r = \sqrt{\sum_{k=1}^{N} z_k^2}$$

[0079]Then, spectral envelope (-) X1340 which a new unit norm spectrum vector had spectral envelope (-) X determined by z/r which broke each slice z by the electric power r, and were normalized as a result are passed to the base extraction process 1360. (-) X means that - is attached on X.

[0080]Spectral envelope (-) X carries out each vector like the line of the form of an observation matrix. The size of a consequential procession is MxN. However, M is the number of time frames and N is the number of frequency (frequency bins). Probably, the procession has the following structures.

[0081]

[Equation 10]

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1^T \\ \tilde{\mathbf{x}}_2^T \\ \vdots \\ \vdots \\ \tilde{\mathbf{x}}_M^T \end{bmatrix}$$

[0082]A base extraction basis function is extracted using singular-value-decomposition SVD130 of drawing 1. SVD is performed using U, S, and command [V] =SVD (X, 0). It is preferred to use "brief" SVD. Brief SVD omits an unnecessary line and a sequence during factorization of SVD. In this invention, since a basis function of a line is unnecessary, extraction efficiency of SVD becomes high. SVD factors a procession as follows. (-) $X=USV^T$, however (-) X are decomposed into a procession product of three processions, U is the Gyoki bottom, S is a diagonal singular-value procession, and V is the transposed sequence basis function. The base is reduced by holding only the first K sequences of only the first K basis functions, i.e., V.

[0083]

[Equation 11]

$$\mathbf{V}_K = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{bmatrix}$$

[0084]However, typically, in the case of an application gestalt by the feature of a sound, there is

K at the range of a basis function of 3-10. In order to determine a rate of information held for K basis functions, a singular value contained in the procession S is used.
[0085]
[Equation 12]

$$I(K)=\frac{\sum_{i=1}^{K} S(i,i)}{\sum_{j=1}^{N} S(j,j)}$$

[0086]However, I (K) is a rate of the information which is held in the case of K basis functions, and N is the total of a basis function equal also to the number of spectra (spectral bins). A SVD basis function is stored in the sequence of the procession.
[0087]In order to give compatibility to the maximum between application gestalten, the function makes the information on k dimension the maximum to other basis functions which can be taken including the sequence in which a basis function has the unit L2 norm. A basis function can be made into a non-orthogonality which is given by an orthogonality which is given by PCA extraction, or ICA extraction. Please refer to the following. Basic projection and reconstruction are described by the following analysis-compositing expressions.
[0088]
[Equation 13]

$$\mathbf{Y}=\mathbf{XV} \qquad (1)$$

$$\mathbf{X}=\mathbf{YV}^{+} \qquad (2)$$

[0089]However, X is a spectral envelope, Y is the spectrum feature, and V is the time feature. The Specht feature is extracted from the mxk observation matrix of the feature, X is a spectrum data matrix of mxn of which the spectrum vector was composed as a line, and V is a nxk procession of the basis function composed by the sequence.
[0090]The first formula corresponds to feature extraction and the 2nd formula corresponds to spectrum reconstruction. Please refer to drawing 8. However, V^{+} expresses the false inverse matrix of V in a non-orthogonality.
[0091]After the SVD base V by which independent-component-analysis reduction was carried out is extracted, the step of an option can perform a basal turn in the independent direction statistically to the maximum extent. This is useful about all the application gestalten which separate the independent component of a spectrogram and need separation of the maximum of the feature. In order to find out the base which became independent statistically using the basis function obtained previously, it is known well and the arbitrary things of the independent-component-analysis (ICA) processes currently introduced broadly can be used. For example, there is JADE or FastICA, Cardoso, J.F. and Laheld, and "Equivariant adaptivesource separation" (IEEE Trans. --) by B.H. [On Signal Processing and] 4: 112 - It is based on 114, 1996 or Hyvarinen, and A ". [Fast and robust fixed-point algorithms for independent component] Please refer to analysis" (IEEE Trans. On Neural Networks, 10(3):626- 634, 1999).
[0092]Use of the following ICA is decomposed into (-) V^{T}_{k} and vector [A] =ica (V^{T}_{k}) which became independent statistically about 1 set of vectors. However, a new base is acquired as a product of a SVD input vector and a false inverse matrix of the presumed mixed procession A which is given by an ICA process. An ICA base is the same size as a SVD base, and is stored in a sequence of a base procession. When the ratio I (K) of information held uses a given extraction

method, it is equivalent to SVD. Basis function(-) V_K1361 is storable in the database 1200. (-) V means that - is attached on V.

[0093]When an input acoustic signal is a mixture generated from many sound sources, clustering of the group of the feature generated by SVD can be carried out as a group by arbitrary known clustering techniques which have a number of dimension equal to a number of dimension of the feature. Thereby, the similar features are collected as the same group. Therefore, each group includes the feature of an audible signal generated by one sound source. The number of groups which will be used in clustering can be set up hand control or automatically according to a level of desired discrimination.

[0094]In order to search for use projection or the time feature Y of a spectrum subspace basis function, the spectral envelope procession X is multiplied by a base vector of the spectrum feature V. This step is the same as any [of SVD and an ICA basis function] case, namely, is (-) Y_k=(-)X(-) V_k. However, Y is a procession which consists of the feature which had a number of dimension after projection of a spectrum over the base V reduced.

[0095]For independent spectrogram reconstruction and visualization, this invention extracts spectrum projection which is not normalized by omitting normalization step 1330 extraction. That is, it is Y_k=X(-) V_k. Since an independent spectrogram is reconstructed here, X_k ingredient as shown in drawing 8, An individual vector pair corresponding to Kth projective vector y_k and Kth reverse base vector v_k is used, and reconstruction type X_k=y_k(-) v^+_k is applied. However, the "+" operator shows transposition for a SVD basis function, and it is a false inverse matrix in the case of whether a SVD basis function is an orthogonality and ICA, and is a non-orthogonality.

[0096]formation of a spectrogram abstract by an independent component -- one of the using forms for a descriptor of these is expressing a spectrogram efficiently by data less than a perfect spectrogram. If an independent component base is used, generally each spectrogram reconstruction thing as shown, for example in drawing 8 corresponds to a sound-source subject in a spectrogram.

[0097]Spending the great portion of difficult work at the time of designing model acquisition and a training sound sorter in collecting and preparing training data. The range of a sound will reflect the range of a category of a sound. For example, a dog barks, each can bark, voice and continuous a large number can bark, and the voice can contain voice or voice at which many dogs bark at once. A model extraction process is adapted for the range of data, and a specimen of a thereby more narrow range generates a sorter specialized more.

[0098]Drawing 14 shows the process 1400 for extracting the feature 1410 and the basis function 1420 as mentioned above from an audible signal generated by the known sound source 1401. Then, Hidden Markov Model is trained using these (1440). A trained model is stored in the database 1200 with those corresponding features. During training, a feature space of n dimension is divided into k states using a clustering process which is not supervised. A feature space is occupied by observation vector which had a number of dimension reduced. The process determines the optimal number of states in the case of given data by reducing a transition matrix, when giving a guess of the first stage for k. Typically, as good sorter performance, five to 10 state is enough.

[0099]Hidden Markov Model is trained in a modification process of a Baum-Welch process which is known also as a Forward-Backward process and which is known well. These processes are extended by implementation of use of a-priori entropy (entropic prior), and deterministic annealing of an expected maximum (EM) process.

[0100]About details about the suitable HMM training process 1430. "It is based on Brand Pattern

discovery. via. entropy. minimization." (Proceedings,Uncertainty'99. Society of Artificial intelligence and Statistics #7, MorganKaufmann, 1999). And "Structure discovery in conditional probability models via an entropic prior and parameter extinction" by Brand. It is indicated to (Neural Computation and 1999).

[0101]After each HMM for a sound source of each known is trained, the model is kept by the permanent memory storage 1200 with a group of the basis function, i.e., the feature of a sound. When a model of many sounds is trained corresponding to the whole classification of a category of a sound, HMM is brought together in a both more big sound recognition sorter data structure, and ontology of a model as shown in drawing 12 by that cause is generated. Indexation is carried out [sound / which is qualitative and has a quantitative description child / new] using the ontology.

[0102]Sound descriptor drawing 15 shows the automatic extracting system 1500 for carrying out [sound / in a database] indexation using a sorter which is kept as a DDL file and which was trained beforehand. A strange sound is read from medium sound-source form like WAV file 1501. Spectrum projection is carried out [sound / the / strange] as mentioned above (1520). Then, one of HMM(s) is chosen from the database 1200 using a group of the projection, i.e., the feature, (1530). Using the Viterbi decoder 1540, it can let a model for the strange sound pass, and both optimal model and a state path can be given. That is, one model state exists to each frame to which windowing was carried out [sound / the]. Please refer to drawing 11 b. Then, indexation of each sound is carried out with the category, model reference, and a model state path, and the descriptor is written in a database in DDL form. Then, it can look for the database 1599 by which indexation was carried out in order to find out a sound where arbitrary descriptors of the above descriptors stored, for example, all the dogs, bark and which is in agreement using voice. Then, an in general similar sound can be provided in the result list 1560.

[0103]Tori's cry, applause, and a dog bark, respectively and, as for drawing 16, classification performance for the classes 1601-1610 of ten sounds, voice, an explosion, a footstep and a sound into which a glass is broken, a report of a gun, sports shoes, a laughing voice, and a telephone is shown. The performance of the system was measured to ground truth using a label of sound effects which are specified by a specialist's sound-effects library. A result shown is a thing for a new sound which is not used during training of a sorter, and, so, illustrates generalization performance of a sorter. The average performance is exact about 95%.

[0104]A section below a specimen search application gestalt gives an example of how to use the recording mode, in order to perform search using both collation by DDL, and reference of medium sound-source form.

[0105]As shown in drawing 17 in a form using DDL by which illustration reference simplification was carried out, the system 1700 is shown reference of a sound using the sound-models state path description 1710 of DDL form. The system reads the reference and occupies in-house-data structure by the descriptive information. This description is compared with description taken out from the database 1599 of a sound stored on a disk (1550). The list 1560 is returned as a result of sorting a best alike sound.

[0106]The sum (SSE) of a square error between state path histograms can be used for the collating step 1550. This collation procedure hardly needs calculation, but can be directly calculated from a state path descriptor stored.

[0107]A certain sound breaks full time length which spends in each state by an overall length of the sound, and a state path histogram gives a discrete frequency function which has a state index as a random variable by that cause. SSE between a reference sound histogram and a histogram of

each sound in a database is used as a range measurement standard. It is more greatly different things that distance is 0, when it suggests that they are the completely same things and distance increases with values other than zero. Using this range measurement standard, a sound in a database is ranked for similarity and a desired number of things are returned as a list in which the nearest thing was first published from a top in that case.
[0108]Drawing 18 a shows a state path and drawing 18 b shows a state path histogram about reference of a sound of a laughing voice. Drawing 19 a shows a state path and drawing 19 b shows a histogram about five sounds which are best in agreement to the reference. All the sounds in agreement are the things from the same class as the reference, and direct that the system is operating correctly.
[0109]In order to use an ontological structure, a sound in an equivalent or category narrower than it which is defined by classification is returned as a sound in agreement. In this way, a "dog" category will return a sound belonging to all the categories related with a "dog" in a certain classification.
[0110]Illustration reference using an audio frequency belt sound and its system can also perform reference which uses an audio frequency belt signal as an input. Here, an input to an illustration reference application gestalt is reference according to an audio frequency belt sound instead of reference by DDL description. In this case, an audio frequency belt sound feature extraction process is performed first, namely, a spectrogram and envelope extraction are performed, and after that, when it is each model in that sorter, projection over a group of a basis function stored is performed.
[0111]The feature which had a number of dimension generated as a result reduced is passed to the Viterbi decoder for a given sorter, and HMM which has the maximum ** score for the given feature is chosen. The Viterbi decoder functions as a model collation algorithm for the system of classification in general. Model reference and a state path are recorded and the result is compared to a database like [in the case of the first example] calculated beforehand.
[0112]It should understand that other various conformity and change may be made by pneuma of this invention and within the limits. So, the purpose of an attached claim is to cover all the modification and change which go into true pneuma of this invention, and within the limits and which start.

[Translation done.]

TECHNICAL FIELD

[Field of the Invention]This invention relates to the method of recognizing and carrying out indexation of the audible signal, and searching for it, in detail about the field of acoustic signal processing generally.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]About extracting the feature of an environmental sound and an ambient sound until now, research was hardly made. The audible signal style of a great portion of conventional technology has been concentrated on human being's sound and music. There is no style for which it was suitable for many sound effects it is heard on the other hand in a movie, television, a video game, and virtual environment, such as umbrella umbrella ******** and a sound into which water flows, that a footstep, a traffic sound, the sound that shuts a door with BATAN, the sound of a laser gun, the sound struck hard, the sound to throw, a thunder, and a leaf are. Generally it is dramatically difficult for these environmental audible signals to extract the feature compared with a sound and music. Because, when those signals are many, it is because the ingredient of piled-up a large number and a high order structural ingredient like repetition and dispersion are included including noise.
[0003]One specific application gestalt which can use such an expression method is graphic processing. The method for extracting, compressing, looking for and classifying an image object is available. For example, please refer to various MPEG standards. Except when the subject of an "audio frequency belt sound" is a sound, the method of processing the subject of such an "audio frequency belt sound" does not exist. For example, while John Wayne shoots 6 running fire handguns, in order to identify the position of all the images along which he is running at full speed on the horse, to search an image library may be desired. To be sure, it is possible to specify John Wayne or a horse visually. However, it is dramatically difficult to sort out the sound of rhythmical PAKAPPAKATSU of a horse which runs at full speed, and the intermittent percussion sound of a revolver. By recognizing the phenomenon of an audio frequency belt sound, the operation within an image can be described in detail.
[0004]Another application gestalt which can use the style is composition of a sound. In order to

compound a sound and to be able to generate by methods other than trial and error, the feature of a sound must be specified before that.
[0005]In conventional technology, the expression for sounds other than a sound reproduces the tone of a specific musical instrument general, for example, It has concentrated on sounds other than a sound of a specific class like identifying the specific musical instrument, distinguishing the sound of a submarine from the sound of the surrounding sea, and recognizing the underwater mammals by the cry. These application gestalten need the acoustical feature of the specific arrangement which is not generalized more than a specific application gestalt, respectively.
[0006]In addition to these specific application gestalten, other researches have been concentrated on developing analysis expression of the acoustical sight generalized. This research came to be known as "auditory scene analysis by calculation." These systems originate in the complexity of the algorithm, and need great computation work. Typically, a method [heuristics / from artificial intelligence and various reasoning methods] is used for those systems.
[0007]Although such a system gives the useful discernment to the arduous problem about sound expression, it is not once shown about a classification and composition of the audible signal in the state where the performance of the system was mixed that it is satisfactory.
[0008]In another application gestalt, indexation of the medium of an audio frequency belt sound including the phenomenon of the sound of the broad range containing cries and music other than an environmental sound, a background noise, sound effects (sound effect), an animal sound, a sound, and a sound can be carried out using expression of a sound. The sound recognition tool for looking for the medium of an audio frequency belt sound can be designed now using the index extracted automatically by this. Semantic description of the contents or the similarity to target reference of an audio frequency belt sound can search for a sound track including the contents of many like a movie or a report program using these tools. For example, to pinpoint the position of all the image clips with which a lion barks or an image raises a cry is desired.
[0009]There is approach in which much realization to automatic classification and indexation is possible. Wold (IEEE Multimedia, pp.27- 36, 1996) etc., "Martin etc. Musical instrument. identificationa pattern-recognition approach: " (Presented at the 136 th Meeting of the Acoustic Society.) of America, Norfolk, VA, and 1998 indicate the strict classification for a musical instrument. "Zhang etc. Content-based. classification and retrieval. of. audio." (SPIE 43rd Annual Meeting, Conference on Advanced Signal Processing Algorithms, Architectures and ImplementationsVIII,.) 1998 indicates the system which trains a model using spectrogram data, "Boreczky etc. A hidden. Markov model framework. for video segmentation using audio and image features" (Proceedings of ICASSP'98, pp.3741- 3744, 1998) uses a Markov model.

[Translation done.]

* NOTICES *

precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]Carrying out indexation of the audio frequency belt sound medium, and looking for it relates to MPEG-7 standards which newly because of multimedia appeared especially closely. The standards need the interface integrated to the class of a general sound. The compatibility of an encoder is one element about a design. In that case, the database of the "sound" which has an index provided by one embodiment can be compared with the database extracted by different embodiment.

[Translation done.]

MEANS

[Means for Solving the Problem]By a computerized method, the feature is extracted from an audible signal generated from one or more sound sources. Windowing of the audible signal is carried out first, it is filtered, and a spectral envelope to each sound source is generated. Then, a number of dimension of a spectral envelope is reduced and the feature which is 1 set for the audible signal is generated. a group [as opposed to / clustering of the feature in the group is carried out, and / each sound source] -- the feature is generated. The feature in each group includes the spectrum feature by which each sound source is characterized, and the corresponding time feature.
[0012]The feature of each group is a quantitative description child, and a quantitative description child is related also with a qualitative descriptor. Hidden Markov Model is trained by group of the known feature, and is stored in a database. Indexation can be carried out by group of the strange feature in order that the database may choose or recognize a similar audible signal in that case.
[0013]

[Embodiment of the Invention]Drawing 1 shows the method 100 for extracting the spectrum feature 108 and the time feature 109 from the mixture 101 of a signal by this invention. The method 100 of this invention is the purpose of classifying a sound source, and it can be used in order to clarify the feature from the recorded sound and to extract it, and in order to change and reuse the purpose in the application gestalt of structurized multimedia like composition of a parameter (re-purpose). The method can be used also in order to extract the feature from the mixture of many dimensions further, the mixture of other linearity, and. The mixture is obtained from one sound source or the sound source of a large number like stereo sound sources.
[0014]In order to extract the feature from the recorded signal, the method by this invention uses the statistical technique based on independent component analysis (ICA). ICA conversion generates rotation of the base of the time-frequency observation matrix 121 using the contrast function defined by the cumulative extension up to the 4th order.
[0015]The base ingredient generated as a result is statistically [as possible] independent, and clarifies the each feature within the mixture sound source 101, for example, the feature of the structure of a sound. The new signal which classifies a signal or has the feature which can be predicted can be specified using such characteristic structures.
[0016]The expression by this invention can compound behavior of many sounds from the feature of a small group. The expression by this invention can compound the characteristic of acoustical subjects, such as complicated acoustical phenomenon structures of colliding, such as bouncing, striking, and grinding, and material, a size, and shape.
[0017]In the method 100, the audio frequency belt sound mixture 101 is first processed by the bank 110 of a logarithmic filter. Each filter generates the band pass signal 111 for a predetermined frequency range. Typically, the band pass signal 111 of 40-50 is generated, and in a low-pass frequency range, many signals are generated from the frequency range of a high region so that the frequency response characteristic of human being's ear may be imitated. In an exception method, as the filter, constant Q (CQ) or a wavelet filter bank can be used, or the filter can be arranged like [in the short-time Fast Fourier Transform expression (STFT)] at linearity.
[0018]In Step 120, for example, each band pass signal is short, "windowing" is carried out to 20msec segment, and an observation matrix is generated. Each procession can also contain hundreds of samples. The details of Steps 110 and 120 are shown in drawing 2 and drawing 3 still in detail. Please care about that windowing can be performed before filtering.
[0019]In Step 130, singular value decomposition (SVD) is applied to the observation matrix 121, and the procession 131 which had the number of dimension reduced is generated. SVD was indicated for the first time by geometrician belt RAMI of Italy in 1873. Singular value decomposition is clear generalization of principal component analysis (PCA). mxn singular-values-of-a-matrix decomposition is arbitrary factorization of the following forms.
[0020]X=UsigmaV ^{T}[0021]However, U has an orthogonal matrix of mxm, U has a regular rectangular cross sequence, V is an orthogonal matrix of nxn, and sigma is a diagonal matrix of mxn of the singular value in which i has ingredient sigma$_{ij}$=0 when not equal to j.
[0022]As one advantage, in contrast with PCA, SVD can decompose a non-square matrix and, thereby, an observation matrix can be directly decomposed in a spectrum or the direction of one of time, without needing to calculate a covariance matrix. Since SVD decomposes a non-square matrix directly, without needing to ask for a covariance matrix, the base generated as a result cannot be easily influenced by PCA to the problem of a dynamic range.
[0023]The method of this invention applies the independent component analysis (ICA) of an option to the procession 131 which had the number of dimension reduced in Step 140. ICA using

the repetitive on-line algorithm based on the false neuro-architecture for blind signal separation is known well. Much neural network architecture for solving an ICA problem these days is proposed. "For example, it was given to Sejnowski on January 17, 1995 Adaptive system for broadband multisignal discrimination in a channel. Please refer to U.S. Pat. No. 5,383,164 of a title called with reverberation."
[0024]ICA generates the spectrum feature 108 and the time feature 109. The spectrum feature expressed as a vector corresponds to the point estimate of the ingredient which is independently statistically in a segmentation window. Similarly the time feature is expressed as a vector and describes deployment of the spectral component in the process of the segment.
[0025]It is combined using the outer product of a vector and each set of a spectrum and a time vector can reconstruct the partial spectrum for a given input spectrum. When reversible, these spectra can presume the independent segment-of-time signal, so that filter bank expression may be so. In the case of the each independent ingredient described in the method, the procession of the compatibility score for the ingredient in a former segment is available. An ingredient can be pursued now over time by presuming the continuous high correspondence of possibility thereby most. Only while seeing the front in time, it is equal to a backward interchangeability procession.
[0026]Each signal component in an audio frequency belt sound track can be presumed using independent component disassembly of the track of an audio frequency belt sound. When the signal procession (mixture of N linearity of N sound sources) of the number of the whole floor cannot be used, it is hard to deal with a separation problem, but the approximation to the sound source in the bottom can be given by using the independent component of the short time section of frequency domain expression. These approximation can be used for a classification, recognition operation, and comparison between sounds.
[0027]As shown in drawing 3, temporal modulation distribution (TFD) can be normalized with the power spectral density (PSD) 115, in order to make small contribution of a lower frequency component which conveys more energies in some sound fields.
[0028]Drawing 4 and drawing 5 show the time and spatial decomposition about the percussion instrument played in a regular rhythm, respectively. By the structure which can be observed, the articulate ingredient of the broadband corresponding to a shake and the horizontal layer system corresponding to the singing of a metal shell become clear.
[0029]Application gestalt this invention for the acoustical feature of a sound can be used in many application gestalten. It can be considered that the extracted feature is a disengageable ingredient showing a peculiar structure in a sound-source mixture of an acoustical mixture. Since the ingredient is recognized or specified, the extracted feature can be compared with 1 set of transcendental classes determined by pattern recognition art. These sorters can be in the field of the analytical model by a phoneme, sound effects, a musical instrument, the animal sound, or other arbitrary corpora. The extracted feature can be individually re-compounded using a reverse filter bank, and, thereby, can attain "purification" of the acoustical mixture of a sound source. The use of an example is separating a singer, a drum, and a guitar from the recorded sound, in order to change the purpose and to reuse some ingredients, or in order to analyze a musical structure automatically. Another example is separating an actor's voice from a background noise and passing a clear audio signal to a speech recognizer, in order to make title translation of the movie automatically.
[0030]The spectrum feature and the time feature can be individually taken into consideration, in order to identify the various characteristics of the acoustical structure of the subject of each sound in a mixture. The time feature can explain behavior of bouncing, breaking, striking, etc. to

the ability of the spectrum feature to explain material, a size, and the characteristic like shape. In this way, it is distinguishable from that a glass bounces or striking earthenware to strike a glass. The extracted feature can be changed and re-compounded in order to generate the synthetic example with which the sound of the sound source was changed. When an input sound is a phenomenon of one sound including two or more acoustical features, such as striking a glass, each feature can be controlled for re-composition. This is useful because of the application gestalt of the medium by models, such as generating a sound in virtual environment.

[0031]Using indexation, search, and this invention, indexation of the big multimedia database including the sound of the type with which many differ, for example, sound effects, an animal sound, a musical instrument, a sound, the overlapping sounds, an environmental sound, a masculine sound, and a feminine sound can be carried out, and it can also search for it.

[0032]Generally in this context, description of a sound is divided into the qualitative description using the character by two types, i.e., a category label, and the quantitive description using a stochastic model state. A category label provides the qualitative information about the contents of the sound. The description in this form is suitable for the application gestalt of reference in a character like the arbitrary processing tools which use the Internet search engine or an alphabetic field.

[0033]By contrast, the quantitative description child can use including the compact information about the segment of an audio frequency belt sound for numerical evaluation of the similarity of a sound. For example, in video or audio sound recording, a specific musical instrument is discriminable using these descriptors. Qualitative and a quantitative description child suits the application gestalt of illustration reference search of an audio frequency belt sound.

[0034]While segmenting in a class a sound recognition descriptor and the audio frequency belt sound by which recording mode qualitative descriptor sound recording was carried out, to acquire the semantic information related about the contents is desired. For example, by recognizing the scream in an image sound track, fear or danger can be directed and a comedy can be directed by a laughing voice. The sound can direct existence of people and, so, the video segments to which these sounds belong can be used as a candidate at the time of looking for a clip including people. The category and system-of-classification descriptor of a sound provide the means for composing a category concept to the layered structure which makes possible this type of complicated related type searching method.

[0035]the category of a sound -- as shown in <u>drawing 6</u> for the easy classification 600, since the category of a sound is named, the recording mode (DS) is used. the sound at which a dog barks as an example -- as a subcategory -- "-- it barks and has voice"611 -- qualitative -- category label "dog" 610 can be given. Furthermore, "groan" 612 or "******" 613 can be made into the desirable subcategory of a "dog." Although the first two subcategories are associated closely, the 3rd subcategory is a phenomenon of a completely different sound. So, <u>drawing 6</u> shows that four categories are composed by the classification which has "dog" 610 as a root node. Each category has at least one related link 601 to another category in the classification. By initial setting, it is considered that the category accommodated is the category (NC) 601 narrower than the accommodated category. however -- this example -- a "groan" -- "-- although it barks and is synonymous with voice" in general, it defines as a thing not more desirable than it. In order to acquire such a structure, the following relations are defined as a part of recording mode of this invention.

[0036]A category larger than BC- means that the category associated is more common in a meaning than in the accommodated category. A category narrower than NC- means that the

category associated is more restrictive in a meaning than in the accommodated category. US - Since it is more desirable than the present category, the present category and the category with which homonymy is related in general are used. UF - Use of the present category is more preferred than the category with which homonymy is related mostly. RC - The category associated is related with homonymy and the category accommodated although it is not homonymy and a larger or narrower category to some extent.
[0037]The following XML-schema codes show how the qualitative recording mode for the categorization method shown in drawing 6 is illustration-ized using a description definition language (DDL).
[0038]
[Equation 2]

```
<SoundCategory term="1" scheme="DOGS">
  <Label>Dogs</Label>
  <TermRelation term="1.1" scheme="DOGS">
    <Label>Bark</Label>
    <TermRelation term="1.2" scheme="DOGS" type="US">
      <Label>Woof</Label>
    </TermRelation>
  </TermRelation>
  <TermRelation term="1.3" scheme="DOGS">
    <Label>Howl</Label>
  </TermRelation>
</SoundCategory>
```

[0039]Both a category and a method attribute provide the peculiar identifier which can be used in order to refer to the following categories and classifications from a quantitative description method like a stochastic model that are indicated still in detail. A label descriptor gives the significant clever semantic label for each category, and relation descriptors describe the relation in the category of the classification by this invention.
[0040]As shown in system-of-classification drawing 7, a category can be combined with the system of classification 700 with a related link, and more abundant classifications can be created. for example, -- "-- it barks, voice"611 are a subcategory of "dog" 610, and "dog" 610 are a subcategory of "pet" 701. As for it, category "cat" 710 are the same. The cat 710 has category "cry" 711 of a sound, and "sound with which throat is sounded" 712. The following is an example of an easy system of classification for a "pet" containing two categories "dog" and a "cat."
[0041]In order to carry out this system of classification by extending a method defined beforehand, the 2nd method to which a name of a "cat" was given is illustration-ized as follows.
[0042]
[Equation 3]

```
<SoundCategory term=”2” scheme="CATS”>
  <Label>Cats</Label>
  <TermRelation term=”2.1 scheme="CATS”>
    <Label>Meow</Label>
  </TermRelation>
  <TermRelation term=”2.2” scheme=”CATS”>
    <Label>Purr</Label>
  </TermRelation>
</SoundCategory>
```

[0043]In order to combine these categories here, the system of classification called a "pet" is illustration-ized with reference to the method defined beforehand.
[0044]
[Equation 4]

```
<ClassificationScheme term=”0” scheme=”PETS”>
  <Label>Pets</Label>
  <ClassificationSchemeRef scheme=”DOGS”/>
  <ClassificationSchemeRef scheme=”CATS”/>
</ClassificationScheme>
```

[0045]Here, the system of classification called a "pet" contains the category "pet" additional as a route including all the category elements of a "dog" and a "cat." In the case of the application gestalt of character indexation, the above qualitative classifications are enough.
[0046]The following sections indicate the quantitative description child for a classification and indexation who is used with a qualitative descriptor and can form the indexation and searching engine of a perfect sound.
[0047]A quantitative description consonant recognition quantitative description child describes the feature of the audible signal which will be used with a statistical sorter. The sound recognition quantitative description child can use for recognition of the general sound containing sound effects and a musical instrument. the descriptor suggested -- in addition, other arbitrary descriptors defined in the structure of an audio frequency belt sound can be used for a classification.
[0048]The feature most extensively used for the classification of the audio frequency band-spectrum base feature sound is a power-spectrum slice or expression by a spectrum like a frame. Typically, each spectrum slice is a vector of n dimension, and n is the number of spectrum channels and has a channel of the data up to 1024 channels. With logarithmic frequency spectrum which is expressed by the structural description child of an audio frequency belt sound, a number of dimension can be reduced to about 32 channels. So, generally, the feature drawn by the spectrum originates in a high number of dimension, and is incompatible with a stochastic model sorter. The probability sorter operates the best with a number of dimension smaller than ten dimensions.
[0049]So, a basis function of the number of low dimensions generated by the above and the following singular value decomposition (SVD) is preferred. In that case, an audio frequency belt sound spectrum base descriptor is a container for a basis function used in order to project the

spectrum on subspace of a low dimension for which it was suitable for a stochastic model sorter.
[0050]This invention determines a base for each class of a sound, and a subclass. The base acquires the statistical most regular feature of a feature space of a sound. Reduction of a number of dimension is performed by projecting a spectrum vector as mentioned above to a procession of a basis function drawn from data. The number of lines corresponds to the length of a spectrum vector, and a basis function is stored in a sequence of a procession corresponding to the number of basis functions in the number of sequences. Base projection is a procession product of a spectrum and a base vector.
[0051]Spectrogram drawing 8 reconstructed from a basis function shows the spectrogram 800 reconstructed from four basis functions by this invention. The concrete spectrogram is a thing for "pop" music. The left-hand side spectrum base vector 801 is combined with the base projective vector 802 using an outer product of a vector. A procession of an outer product generated as a result, respectively is added, and a final reconstruction thing is generated. A basis function is chosen so that information may be made into the maximum in a small number of dimension from the original data. For example, the basis function can respond to a statistically independent ingredient which corresponds to the feature of not correlating [which is extracted using principal component analysis (PCA) or Karhunen-Loeve conversion (KLT)], or is extracted by independent component analysis (ICA). When a secondary statistic, i.e., covariance, understands KLT or the Hotelling conversion, it is desirable negative correlation conversion. This reconstruction is indicated still in detail with reference to drawing 13.
[0052]In order to achieve the purpose of a classification, a base for all the classes is drawn. In this way, classifying space contains a statistically remarkable ingredient of the class. The following DDL illustration-ization defines a base projection matrix which reduces a series of logarithmic frequency spectrum of 31 channels to five dimensions.
[0053]
[Equation 5]

```
<AudioSpectrumBasis loEdge="62.5" hiEdge="8000" resolution="1/4 octave">
  <Basis>
     <Matrix dim="31 5">
     0.26 -0.05 0.01 -0.70 0.44
     0.34 0.09 0.21 -0.42 -0.05
     0.33 0.15 0.24 -0.05 -0.39
     0.33 0.15 0.24 -0.05 -0.39
     0.27 0.13 0.16 0.24 -0.04
     0.27 0.13 0.16 0.24 -0.04
     0.23 0.13 0.09 0.27 0.24
     0.20 0.13 0.04 0.22 0.40
     0.17 0.11 0.01 0.14 0.37
    ...
        </Matrix>
  </Basis>
</AudioSpectrumBasis>
```

[0054]Low edge, high edge, and a resolution attribute give the interval of the spectrum channel

in the bottom frequency limit, the upper part frequency limit, and octave-band notation of a basis function. In the classification structure by this invention, the basis function for all the classes of a sound is stored with the stochastic model for the class.
[0055]The features used for the feature sound recognition of sound recognition can be collected, and the one recording mode which can be used for various different application gestalten can be adopted. In a classification of the sound acquired, many types, for example, sound-effects libraries, of the sound, and the sample disk of a musical instrument, the audio frequency belt sound spectrum projection descriptor of initial setting plays a role good.
[0056]The base feature is drawn from the above audio frequency belt sound spectral envelope extraction processes. An audio frequency belt sound spectrum projection descriptor is a container for the feature which reduced a number of dimension similarly obtained by the projection of a spectral envelope to 1 set of basis functions as mentioned above. For example, an audio frequency belt sound spectral envelope is extracted by sliding window FFT analysis with re-sampling to a frequency band arranged by logarithm. In a desirable embodiment, analysis frame periods are 10msec. However, a slide extraction window of 30msec temporal duration is used in a Hamming window. 30msec interval provides sufficient spectral resolving power, and it is chosen so that a channel of the beginning of the 62.5-Hz width of an octave-band spectrum may be decomposed in general. A size of an FFT-analysis window is a sample number of involution of 2 big next. As for this, in 30msec, 960 samples exist at 32 kHz, but it means that FFT will be performed in 1024 samples. In 30msec, 1323 samples exist at 44.1 kHz, but FFT will be performed in 2048 samples and a sample outside a window is set as 0.
[0057]Drawing 9 a and drawing 9 b show the base projection 911-913 in the case of the frequency index 920 for the "laughing voice" spectrogram 1000 in the three spectrum base ingredients 901-903, drawing 10 a, and drawing 10 b of a case of the time index 910 generated. A form here is similar to form shown in drawing 4 and drawing 5. Drawing 10 a shows a spectrogram of logarithmic scale of a laughing voice, and drawing 10 b shows what reconstructed a spectrogram. Which drawing plots time and a frequency index on a x axis and the y-axis, respectively.
[0058]In addition to a base descriptor, a sorter can be defined using the special characteristic of a class of a sound like a harmonic envelope for a musical instrument classification used in many cases, and the fundamental frequency feature using another quantitative description child's big sequence.
[0059]The one convenience of number-of-dimension reduction which is made by this invention is that arbitrary descriptors based on 1 set of descriptors in which scaling is possible can add to a spectrum descriptor with the same sampling rate. A suitable base is calculable to the whole group of the extended feature like a base based on a spectrum.
[0060]Another application gestalt for the feature recording mode of sound recognition by spectrogram abstract-ized this invention using a basis function is an efficient spectrogram expression. Visualization and in order to abstract-ize, audio frequency belt sound spectrum base projection and the audio frequency belt sound spectrum base feature can be used for a spectrogram as dramatically efficient memory mechanisms.
[0061]Since a spectrogram is reconstructed, this invention uses the formula 2 indicated in detail by the following. The formula 2 constitutes a two-dimensional spectrogram from a cross product of each basis function shown also in drawing 8 as mentioned above, and its corresponding spectrogram basic projection.
[0062]Since the stochastic model recording mode finite state model spectrum feature is changed

over time, a phenomenon of a sound is dynamic. This very big time change gives characteristic "fingerprint" for recognition to an audible signal. So, a model of this invention divides into a limited state number an audible signal generated by class of a specific sound source or a sound. The division is based on the spectrum feature. Each sound is described by locus passing through this state space of those sounds. This model is indicated in detail by the following about drawing 11 a and drawing 11 b. Each state can be expressed by continuous probability distribution like Gaussian distribution.

[0063]Dynamic behavior of a class of a sound passing through state space is expressed by transition matrix of kxk which describes probability of transition in the following state when giving the present state. The transition matrix T models probability of transition in the state j in the time t from the state i in time t-l. Early state distribution is kx1 vector of probability, and is typically used also in a finite state model. The k-th element of this vector is probability which is in the state k in the first observation frame.

[0064]Since a state is modeled during a classification of a sound, Gaussian distribution type multi-dimension Gaussian distribution is used. Gaussian distribution is parameterized by a 1xn vector of the average value m, and the covariance matrix K of nxn. However, n is the number of the features in each observation vector. When a gauss parameter is given, a formula for calculation of probability to the specific vector x is as follows.

[0065]

[Equation 6]

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \mathbf{K}^{-1} (\mathbf{x} - \mathbf{m}) \right]$$

[0066]Continuation Hidden Markov Model is a finite state model which has a continuous-probability-distribution model for state observation probability. The following DDL illustration-ization is examples of use of the stochastic model recording mode for expressing the continuation Hidden Markov Model which has a gauss state. Below as for the decimal point, in this example, the floating point number is rounded off by double figures only for the purpose of a display.

[0067]

[Equation 7]

```
<ProbabilityModel xsi:type="ContinuousMarkovModelType" numberStates="7">
<Initial dim="7">
0.04 0.34 0.12 0.04 0.34 0.12 0.00 </Initial>
<Transitions dim="7 7">
0.91 0.02 0.00 0.00 0.05 0.01 0.01
0.01 0.99 0.00 0.00 0.00 0.00 0.00
0.01 0.00 0.92 0.01 0.01 0.06 0.00
0.00 0.00 0.00 0.99 0.01 0.00 0.00
0.02 0.00 0.00 0.00 0.97 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.98 0.01
0.02 0.00 0.00 0.00 0.00 0.02 0.96
</Transitions>
<State><Label>1</Label></State>
<!—State 1 Observation Distribution -- >
<ObservationDistribution xsi:type="GaussianDistributionType">
<Mean dim="6">
5.11 -9.28 -0.69 -0.79 0.38 0.47
</Mean>
<Covariance dim="6 6">
1.40 -0.12 -1.53 -0.72 0.09 -1.26
-0.12 0.19 0.02 -0.21 0.23 0.17
-1.53 0.02 2.44 1.41 -0.30 1.69
-0.72 -0.21 1.41 2.27 -0.15 1.05
0.09 0.23 -0.30 -0.15 0.80 0.29
-1.26 0.17 1.69 1.05 0.29 2.24
</Covariance>
<State><Label>2</Label></State>
<!—Remaining states use same structures-- >
<\PobabilityModel>
```

[0068]In this example, a "stochastic model" is illustration-ized as a Gaussian distribution type drawn from a base stochastic model class.
[0069]By the method by this invention, the tool has so far [sound recognized model recording mode] been separated, without completely using the structure of an application gestalt. The following data types combine an above-mentioned descriptor and recording mode, and make them the structure where it was unified for the classification of a sound, and indexation. Indexation of the segment of a sound can be carried out with the category label based on the output of a sorter. A probability model parameter can be used for the indexation of the sound in a database. To carry out indexation by a model parameter like a state is needed by the illustration reference application gestalt, when a reference category is strange, or when a collation judgment standard narrower than the range of a category is needed.
[0070]The sound recognized model sound recognized model recording mode specifies the stochastic model of the class of a sound like Hidden Markov Model or a gauss mixing model. the following examples -- drawing 6 -- "-- barking -- the Hidden Markov Model of the voice" sound category 611 -- **** -- it is-izing. The stochastic model and the related basis function for the class of the sound are similarly defined as the case of the example indicated previously.
[0071]
[Equation 8]

```
<SoundRecognitionModel id="sfx1.1" SoundCategoryRef="Bark">
<ExtractionInformation term="Parameters" scheme="ExtractionParameters">
<Label>NumStates=7, NumBasisComponents=5</Label>
</ExtractionInformation>
<ProbabilityModel xsi:type="ContinuousMarkovModelType" numberStates="7">
    ... <!-- see previous example -->
</ProbabilityModel>
<SpectrumBasis loEdge="62.5" hiEdge="8000" resolution="1/4 octave">
    ... <!-- see previous example -->
</SpectrumBasis>
</SoundRecognitionModel>
```

[0072]sound-models state Pass -- this descriptor describes the dynamic state path of a sound through that model with reference to a finite state stochastic model. Indexation can be carried out [sound] in two modes segmenting a sound in the model state, or by sampling a state path at a regular interval. As for the temporal duration of the segment, in the case of the 1st, each audio frequency belt sound segment directs the effective temporal duration for the state including the reference to one state. In the case of the 2nd, a sound is described by the index of a sampled series which refers to a model state. The sound category which has comparatively long state temporal duration is efficiently described using one segment and 1 state approach. The sound which has comparatively short state temporal duration is described still more efficiently using a series of sampled state indices.
[0073]Drawing 11 a shows the logarithmic spectrogram (frequency versus time) 1100 of the dog **** vocal sound 611 of drawing 6. Drawing 11 b shows a sound-models state path sequence in the state where drawing 11 a barked and it let continuation Hidden Markov Model for a voice model pass covering the same time interval. In drawing 11 b, a x axis is a time index and the y-axis is a state index.
[0074]Sound recognition sorter drawing 12 shows a sound recognition sorter which uses the one database 1200 for all the required ingredients of a sorter. The sound recognition sorter describes a relation between many stochastic models, and, thereby, defines ontology of a sorter. for example, a class of an extensive sound like [so that a hierarchical recognizer may be indicated in the case of drawing 6 and drawing 7] an animal in a root node -- moreover -- Dog: barking in a leaf node -- voice and a cat:cry, and ** -- a finer class [like] can be classified. Using descriptor method structure of a graph, this method defines a correspondence relation between ontology of a sorter, and classification of a category of a sound, and in order that hierarchical sound models may extract category description in the case of given classification, it is used.
[0075]Drawing 13 shows the system 1300 for constituting a database of a model. A system shown in drawing 13 is an extended type of a system shown in drawing 1. Here, before filtering in order to extract a spectral envelope, windowing of the input acoustic signal is carried out. The system can incorporate the audio frequency belt sound inputting 1301 in a form of an audio file of WAV form, for example. The system extracts the audio frequency belt sound feature from a file, and trains Hidden Markov Model in these features. In the case of a class of each sound, a directory of a specimen of a sound is used for the system. Hierarchical directory structure defines ontology corresponding to desired classification. In the case of each directory of the ontology, one Hidden Markov Model is trained.
[0076]The system 1300 of audio frequency belt sound feature extraction drawing 13 shows a method for extracting an audio frequency belt sound spectrum basis function and the feature from an audible signal as mentioned above. The input acoustic signal 1301 is generable by one sound source, for example, a person, animal, a musical instrument, or many sound sources, for

example, people and an animal, many musical instruments or composite tone. In the case of the latter, an audible signal is a mixture. Windowing of the input acoustic signal is first carried out to 10msec frame (1310). In drawing 1, an input signal should care about that band pass filtering is carried out before windowing. Here, windowing of the audible signal is carried out first, it is filtered after that (1320), and extracts frequency spectrum (short-time logarithmic-in-frequency spectrum) logarithmic [short-time]. Filtering performs time-frequency power-spectrum analysis like short time Fourier transformation (squared-magnitude) which squared a size. The result is a procession which has M frames and the frequency (frequency bins) of N pieces. The spectrum vector x is a line of this procession.
[0077]Step 1330 performs normalization of logarithmic scale. Each spectrum vector x is changed into the decibel graduation 1331 from a power spectrum by $z=10\log_{10}(x)$. Step 1332 determines the L2 norm of a vector element as follows.
[0078]
[Equation 9]

$$r = \sqrt{\sum_{k=1}^{N} z_k^2}$$

[0079]Then, spectral envelope (-) X1340 which a new unit norm spectrum vector had spectral envelope (-) X determined by z/r which broke each slice z by the electric power r, and were normalized as a result are passed to the base extraction process 1360. (-) X means that - is attached on X.
[0080]Spectral envelope (-) X carries out each vector like the line of the form of an observation matrix. The size of a consequential procession is MxN. However, M is the number of time frames and N is the number of frequency (frequency bins). Probably, the procession has the following structures.
[0081]
[Equation 10]

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1^T \\ \tilde{\mathbf{x}}_2^T \\ \vdots \\ \vdots \\ \tilde{\mathbf{x}}_M^T \end{bmatrix}$$

[0082]A base extraction basis function is extracted using singular-value-decomposition SVD130 of drawing 1. SVD is performed using U, S, and command [V] =SVD (X, 0). It is preferred to use "brief" SVD. Brief SVD omits an unnecessary line and a sequence during factorization of SVD. In this invention, since a basis function of a line is unnecessary, extraction efficiency of SVD becomes high. SVD factors a procession as follows. (-) $X=USV^T$, however (-) X are decomposed into a procession product of three processions, U is the Gyoki bottom, S is a diagonal singular-value procession, and V is the transposed sequence basis function. The base is reduced by holding only the first K sequences of only the first K basis functions, i.e., V.
[0083]
[Equation 11]

$$\mathbf{V}_K = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{bmatrix}$$

[0084]However, typically, in the case of an application gestalt by the feature of a sound, there is

K at the range of a basis function of 3-10. In order to determine a rate of information held for K basis functions, a singular value contained in the procession S is used.
[0085]
[Equation 12]

$$I(K)=\frac{\sum_{i=1}^{K}S(i,i)}{\sum_{j=1}^{N}S(j,j)}$$

[0086]However, I (K) is a rate of the information which is held in the case of K basis functions, and N is the total of a basis function equal also to the number of spectra (spectral bins). A SVD basis function is stored in the sequence of the procession.
[0087]In order to give compatibility to the maximum between application gestalten, the function makes the information on k dimension the maximum to other basis functions which can be taken including the sequence in which a basis function has the unit L2 norm. A basis function can be made into a non-orthogonality which is given by an orthogonality which is given by PCA extraction, or ICA extraction. Please refer to the following. Basic projection and reconstruction are described by the following analysis-compositing expressions.
[0088]
[Equation 13]

$$\mathbf{Y}=\mathbf{X}\mathbf{V} \qquad (1)$$

$$\mathbf{X}=\mathbf{Y}\mathbf{V}^{+} \qquad (2)$$

[0089]However, X is a spectral envelope, Y is the spectrum feature, and V is the time feature. The Specht feature is extracted from the mxk observation matrix of the feature, X is a spectrum data matrix of mxn of which the spectrum vector was composed as a line, and V is a nxk procession of the basis function composed by the sequence.
[0090]The first formula corresponds to feature extraction and the 2nd formula corresponds to spectrum reconstruction. Please refer to drawing 8. However, V^+ expresses the false inverse matrix of V in a non-orthogonality.
[0091]After the SVD base V by which independent-component-analysis reduction was carried out is extracted, the step of an option can perform a basal turn in the independent direction statistically to the maximum extent. This is useful about all the application gestalten which separate the independent component of a spectrogram and need separation of the maximum of the feature. In order to find out the base which became independent statistically using the basis function obtained previously, it is known well and the arbitrary things of the independent-component-analysis (ICA) processes currently introduced broadly can be used. For example, there is JADE or FastICA, Cardoso, J.F. and Laheld, and "Equivariant adaptivesource separation" (IEEE Trans. --) by B.H. [On Signal Processing and] 4: 112 - It is based on 114, 1996 or Hyvarinen, and A ". [Fast and robust fixed-point algorithms for independent component] Please refer to analysis" (IEEE Trans. On Neural Networks, 10(3):626- 634, 1999).
[0092]Use of the following ICA is decomposed into (-) V^T_k and vector [A] =ica (V^T_k) which became independent statistically about 1 set of vectors. However, a new base is acquired as a product of a SVD input vector and a false inverse matrix of the presumed mixed procession A which is given by an ICA process. An ICA base is the same size as a SVD base, and is stored in a sequence of a base procession. When the ratio I (K) of information held uses a given extraction

method, it is equivalent to SVD. Basis function(-) V_K1361 is storable in the database 1200. (-) V means that - is attached on V.
[0093]When an input acoustic signal is a mixture generated from many sound sources, clustering of the group of the feature generated by SVD can be carried out as a group by arbitrary known clustering techniques which have a number of dimension equal to a number of dimension of the feature. Thereby, the similar features are collected as the same group. Therefore, each group includes the feature of an audible signal generated by one sound source. The number of groups which will be used in clustering can be set up hand control or automatically according to a level of desired discrimination.
[0094]In order to search for use projection or the time feature Y of a spectrum subspace basis function, the spectral envelope procession X is multiplied by a base vector of the spectrum feature V. This step is the same as any [of SVD and an ICA basis function] case, namely, is (-) Y_k=(-)X(-) V_k. However, Y is a procession which consists of the feature which had a number of dimension after projection of a spectrum over the base V reduced.
[0095]For independent spectrogram reconstruction and visualization, this invention extracts spectrum projection which is not normalized by omitting normalization step 1330 extraction. That is, it is Y_k=X(-) V_k. Since an independent spectrogram is reconstructed here, X_k ingredient as shown in drawing 8, An individual vector pair corresponding to Kth projective vector y_k and Kth reverse base vector v_k is used, and reconstruction type X_k=y_k(-) v^+_k is applied. However, the "+" operator shows transposition for a SVD basis function, and it is a false inverse matrix in the case of whether a SVD basis function is an orthogonality and ICA, and is a non-orthogonality.
[0096]formation of a spectrogram abstract by an independent component -- one of the using forms for a descriptor of these is expressing a spectrogram efficiently by data less than a perfect spectrogram. If an independent component base is used, generally each spectrogram reconstruction thing as shown, for example in drawing 8 corresponds to a sound-source subject in a spectrogram.
[0097]Spending the great portion of difficult work at the time of designing model acquisition and a training sound sorter in collecting and preparing training data. The range of a sound will reflect the range of a category of a sound. For example, a dog barks, each can bark, voice and continuous a large number can bark, and the voice can contain voice or voice at which many dogs bark at once. A model extraction process is adapted for the range of data, and a specimen of a thereby more narrow range generates a sorter specialized more.
[0098]Drawing 14 shows the process 1400 for extracting the feature 1410 and the basis function 1420 as mentioned above from an audible signal generated by the known sound source 1401. Then, Hidden Markov Model is trained using these (1440). A trained model is stored in the database 1200 with those corresponding features. During training, a feature space of n dimension is divided into k states using a clustering process which is not supervised. A feature space is occupied by observation vector which had a number of dimension reduced. The process determines the optimal number of states in the case of given data by reducing a transition matrix, when giving a guess of the first stage for k. Typically, as good sorter performance, five to 10 state is enough.
[0099]Hidden Markov Model is trained in a modification process of a Baum-Welch process which is known also as a Forward-Backward process and which is known well. These processes are extended by implementation of use of a-priori entropy (entropic prior), and deterministic annealing of an expected maximum (EM) process.
[0100]About details about the suitable HMM training process 1430. "It is based on Brand Pattern

discovery. via. entropy. minimization." (Proceedings,Uncertainty'99. Society of Artificial intelligence and Statistics #7, MorganKaufmann, 1999). And "Structure discovery in conditional probability models via an entropic prior and parameter extinction" by Brand. It is indicated to (Neural Computation and 1999).

[0101]After each HMM for a sound source of each known is trained, the model is kept by the permanent memory storage 1200 with a group of the basis function, i.e., the feature of a sound. When a model of many sounds is trained corresponding to the whole classification of a category of a sound, HMM is brought together in a both more big sound recognition sorter data structure, and ontology of a model as shown in drawing 12 by that cause is generated. Indexation is carried out [sound / which is qualitative and has a quantitative description child / new] using the ontology.

[0102]Sound descriptor drawing 15 shows the automatic extracting system 1500 for carrying out [sound / in a database] indexation using a sorter which is kept as a DDL file and which was trained beforehand. A strange sound is read from medium sound-source form like WAV file 1501. Spectrum projection is carried out [sound / the / strange] as mentioned above (1520). Then, one of HMM(s) is chosen from the database 1200 using a group of the projection, i.e., the feature, (1530). Using the Viterbi decoder 1540, it can let a model for the strange sound pass, and both optimal model and a state path can be given. That is, one model state exists to each frame to which windowing was carried out [sound / the]. Please refer to drawing 11 b. Then, indexation of each sound is carried out with the category, model reference, and a model state path, and the descriptor is written in a database in DDL form. Then, it can look for the database 1599 by which indexation was carried out in order to find out a sound where arbitrary descriptors of the above descriptors stored, for example, all the dogs, bark and which is in agreement using voice. Then, an in general similar sound can be provided in the result list 1560.

[0103]Tori's cry, applause, and a dog bark, respectively and, as for drawing 16, classification performance for the classes 1601-1610 of ten sounds, voice, an explosion, a footstep and a sound into which a glass is broken, a report of a gun, sports shoes, a laughing voice, and a telephone is shown. The performance of the system was measured to ground truth using a label of sound effects which are specified by a specialist's sound-effects library. A result shown is a thing for a new sound which is not used during training of a sorter, and, so, illustrates generalization performance of a sorter. The average performance is exact about 95%.

[0104]A section below a specimen search application gestalt gives an example of how to use the recording mode, in order to perform search using both collation by DDL, and reference of medium sound-source form.

[0105]As shown in drawing 17 in a form using DDL by which illustration reference simplification was carried out, the system 1700 is shown reference of a sound using the sound-models state path description 1710 of DDL form. The system reads the reference and occupies in-house-data structure by the descriptive information. This description is compared with description taken out from the database 1599 of a sound stored on a disk (1550). The list 1560 is returned as a result of sorting a best alike sound.

[0106]The sum (SSE) of a square error between state path histograms can be used for the collating step 1550. This collation procedure hardly needs calculation, but can be directly calculated from a state path descriptor stored.

[0107]A certain sound breaks full time length which spends in each state by an overall length of the sound, and a state path histogram gives a discrete frequency function which has a state index as a random variable by that cause. SSE between a reference sound histogram and a histogram of

each sound in a database is used as a range measurement standard. It is more greatly different things that distance is 0, when it suggests that they are the completely same things and distance increases with values other than zero. Using this range measurement standard, a sound in a database is ranked for similarity and a desired number of things are returned as a list in which the nearest thing was first published from a top in that case.
[0108]Drawing 18 a shows a state path and drawing 18 b shows a state path histogram about reference of a sound of a laughing voice. Drawing 19 a shows a state path and drawing 19 b shows a histogram about five sounds which are best in agreement to the reference. All the sounds in agreement are the things from the same class as the reference, and direct that the system is operating correctly.
[0109]In order to use an ontological structure, a sound in an equivalent or category narrower than it which is defined by classification is returned as a sound in agreement. In this way, a "dog" category will return a sound belonging to all the categories related with a "dog" in a certain classification.
[0110]Illustration reference using an audio frequency belt sound and its system can also perform reference which uses an audio frequency belt signal as an input. Here, an input to an illustration reference application gestalt is reference according to an audio frequency belt sound instead of reference by DDL description. In this case, an audio frequency belt sound feature extraction process is performed first, namely, a spectrogram and envelope extraction are performed, and after that, when it is each model in that sorter, projection over a group of a basis function stored is performed.
[0111]The feature which had a number of dimension generated as a result reduced is passed to the Viterbi decoder for a given sorter, and HMM which has the maximum ** score for the given feature is chosen. The Viterbi decoder functions as a model collation algorithm for the system of classification in general. Model reference and a state path are recorded and the result is compared to a database like [in the case of the first example] calculated beforehand.
[0112]It should understand that other various conformity and change may be made by pneuma of this invention and within the limits. So, the purpose of an attached claim is to cover all the modification and change which go into true pneuma of this invention, and within the limits and which start.

[Translation done.]

DESCRIPTION OF DRAWINGS

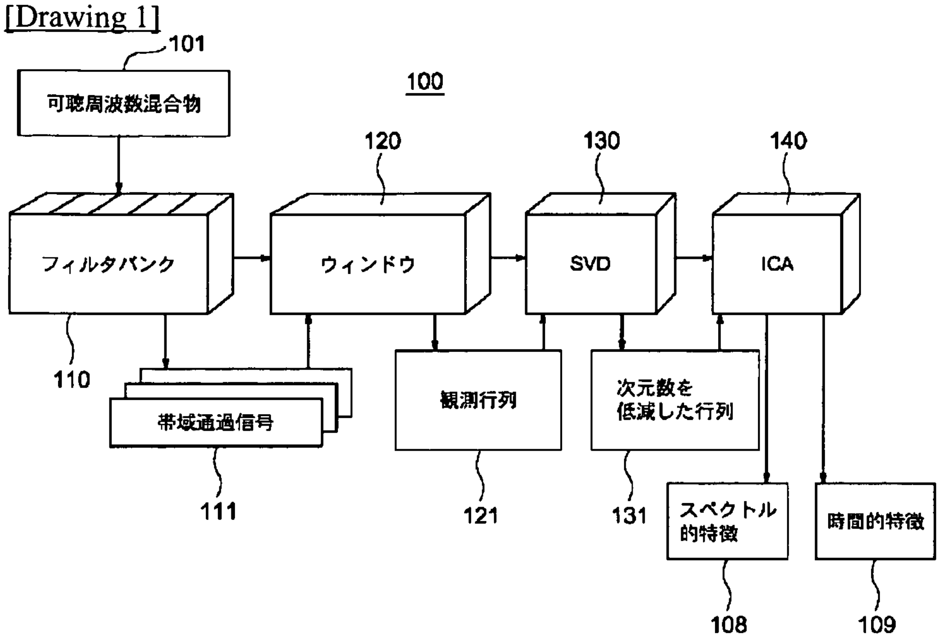
[Brief Description of the Drawings]
[Drawing 1]It is a flow chart of the method for extracting the feature from the mixture of the signal by this invention.
[Drawing 2]It is a block diagram of filtering and a window processing step.
[Drawing 3]It is a block diagram of the step normalized, reduced and extracted.
[Drawing 4]It is a graph of the feature of a metal percussion instrument.
[Drawing 5]It is a graph of the feature of a metal percussion instrument.
[Drawing 6]It is a block diagram of the verbal model about the voice at which a dog barks.
[Drawing 7]It is a block diagram of the verbal model about a pet's sound.
[Drawing 8]It is a spectrogram reconstructed from four spectrum basis functions and base projection.
[Drawing 9 a] It is a base projection envelope about a laughing voice.
[Drawing 9 b] It is an audio frequency belt sound spectrum about the laughing voice of drawing 9 a.
[Drawing 10 a] It is a spectrogram of the logarithmic scale about a laughing voice.
[Drawing 10 b] It is the reconstructed spectrogram about a laughing voice.
[Drawing 11 a] It is a spectrogram of logarithmic scale in case a dog barks.
[Drawing 11 b] It is a sequence diagram of the sound-models state path in the state where it let continuation Hidden Markov Model in case the dog of drawing 11 a barks pass.
[Drawing 12]It is a block diagram of a sound recognition sorter.
[Drawing 13]It is a block diagram of the system for extracting the sound by this invention.
[Drawing 14]It is a block diagram of the process for training the Hidden Markov Model by this invention.
[Drawing 15]It is a block diagram of the system for specifying and classifying the sound by this invention.
[Drawing 16]It is a graph of the performance of the system of drawing 15.
[Drawing 17]It is a block diagram of the sound inquiry system by this invention.
[Drawing 18 a] It is a block diagram of the state path of a laughing voice.
[Drawing 18 b] It is a histogram of the state path of a laughing voice.
[Drawing 19 a] It is a figure showing the state path of a laughing voice in agreement.
[Drawing 19 b] It is a histogram of the state path of a laughing voice in agreement.

[Translation done.]

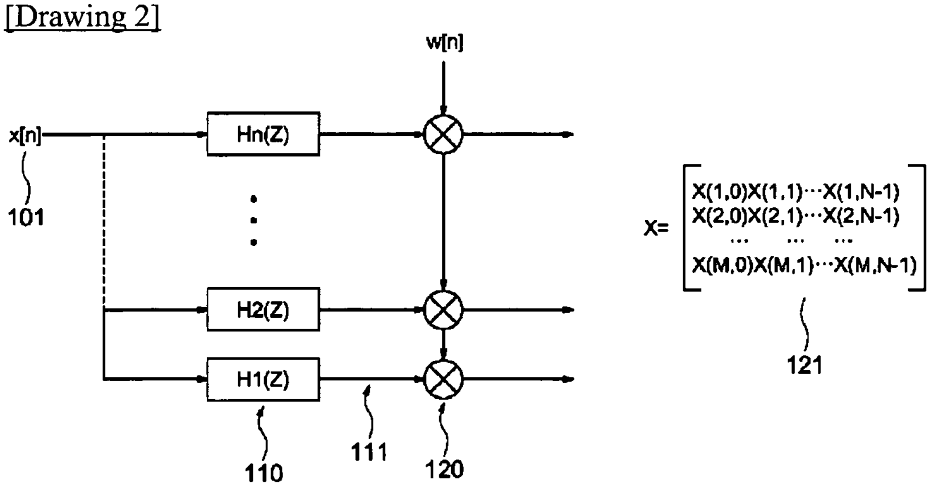
precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

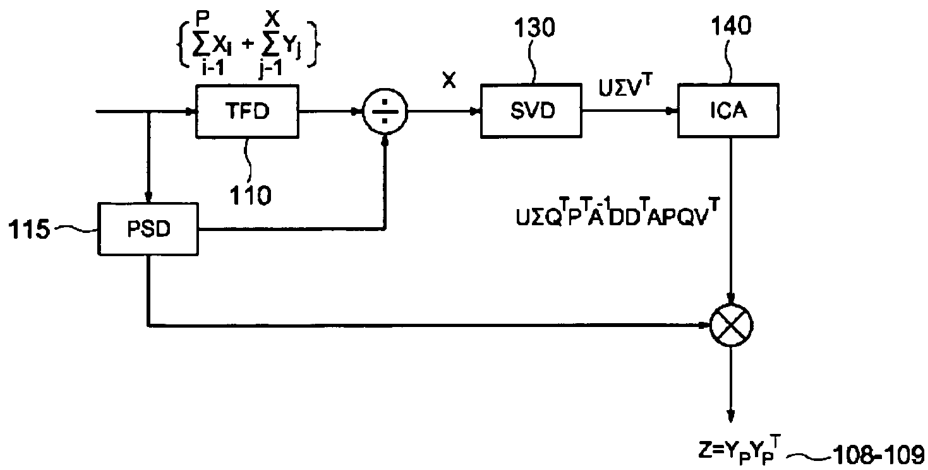
[Drawing 1]



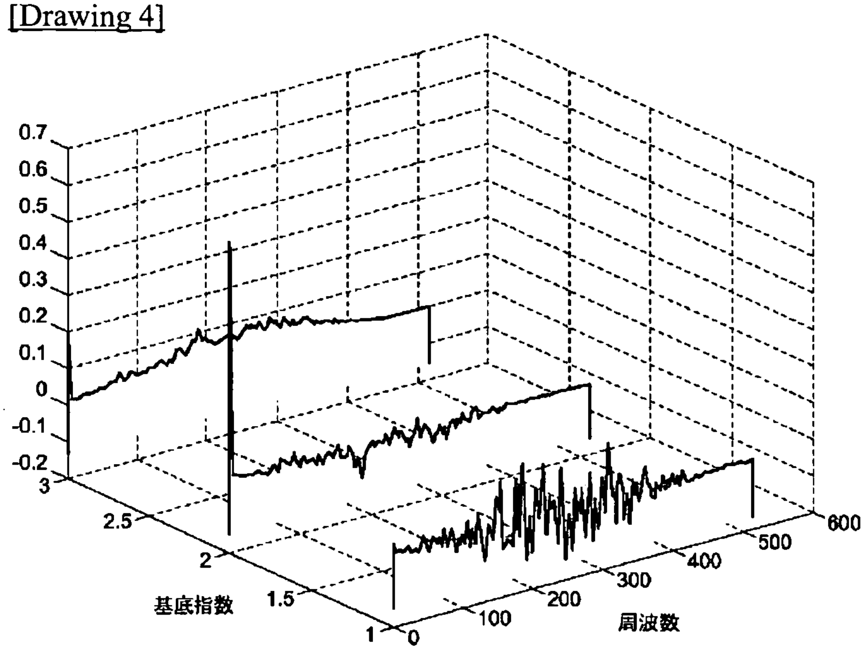
[Drawing 2]



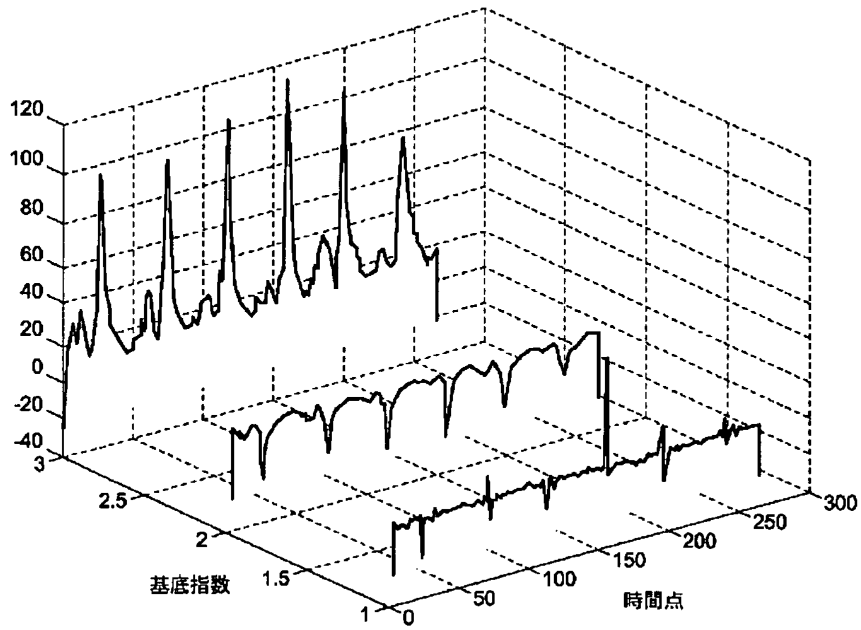
[Drawing 3]



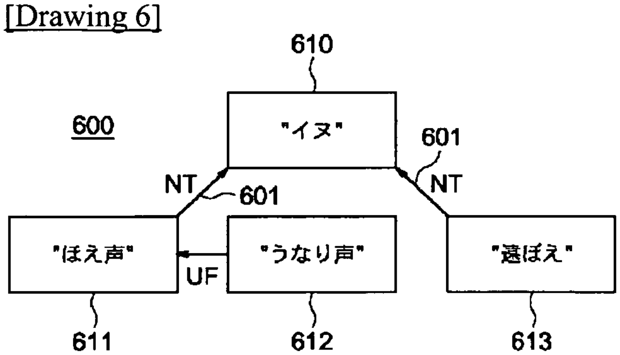
[Drawing 4]



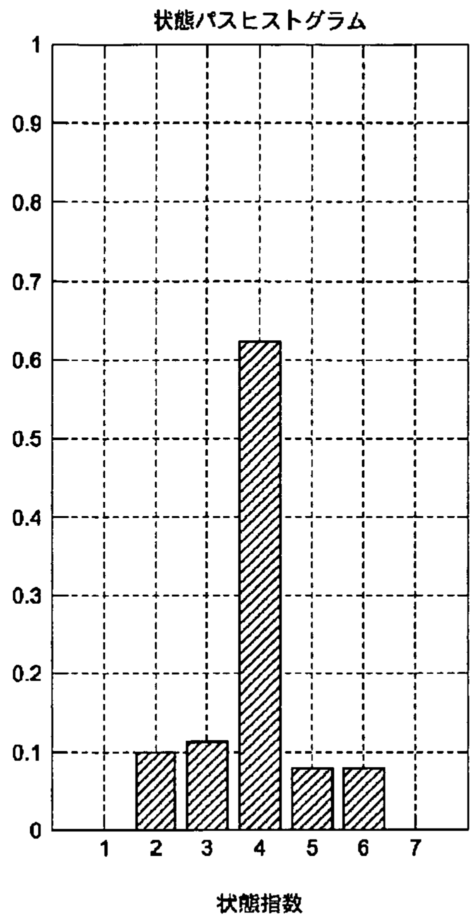
[Drawing 5]



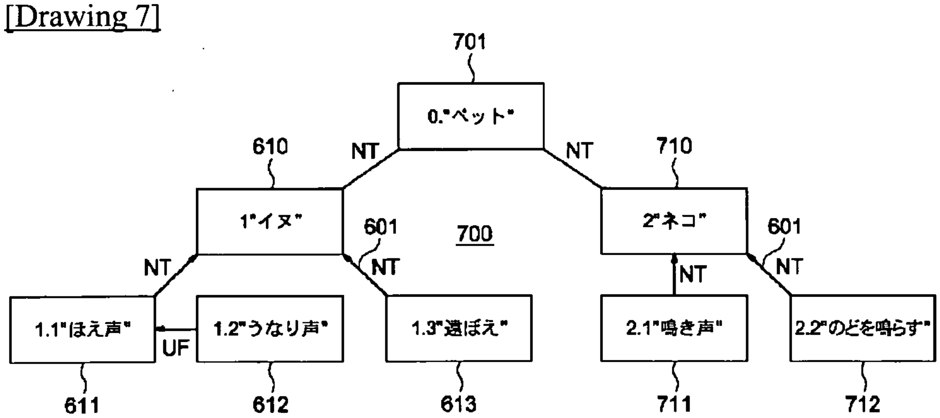
[Drawing 6]



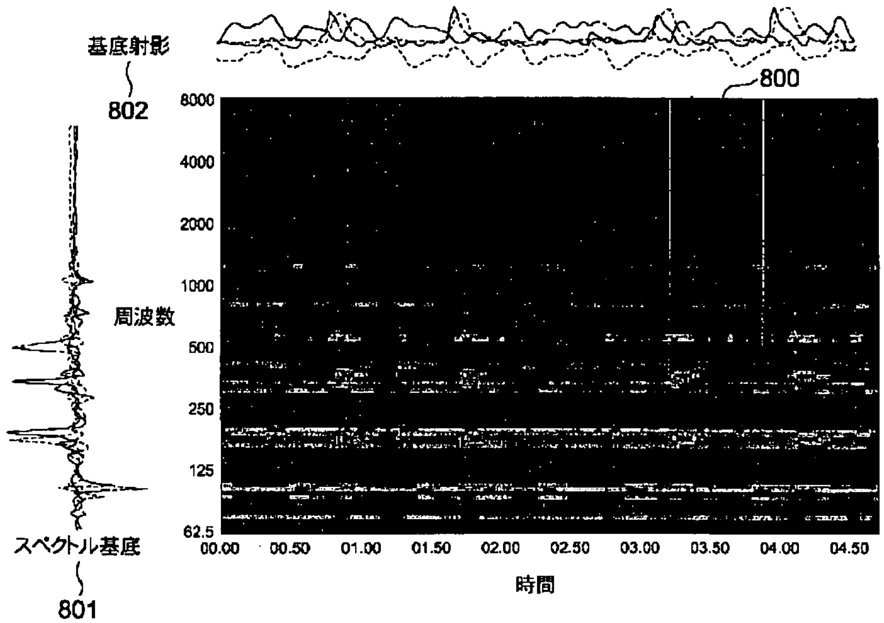
[Drawing 18 b]



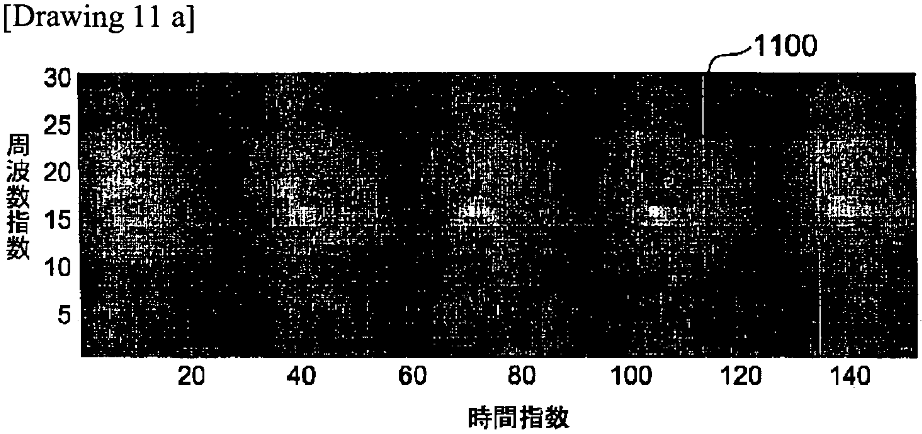
[Drawing 7]



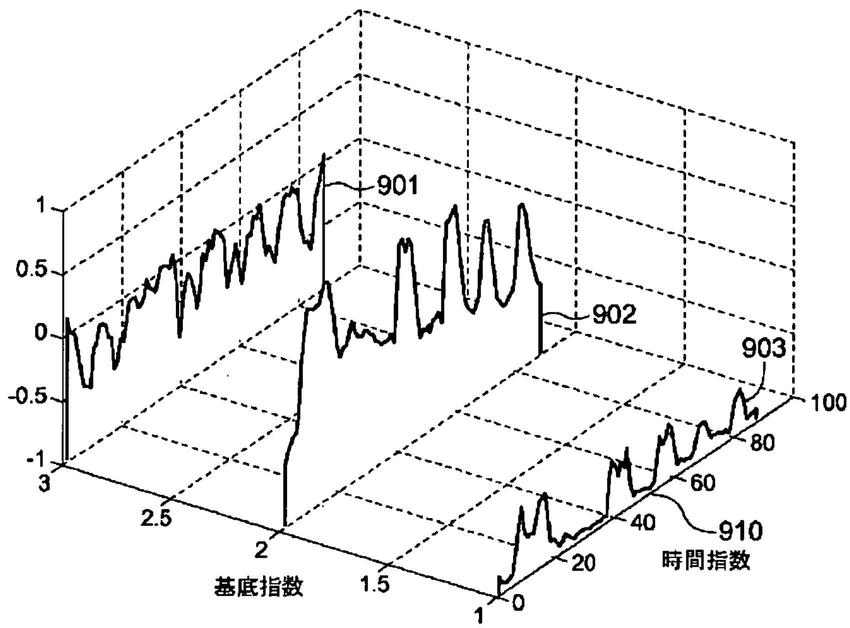
[Drawing 8]



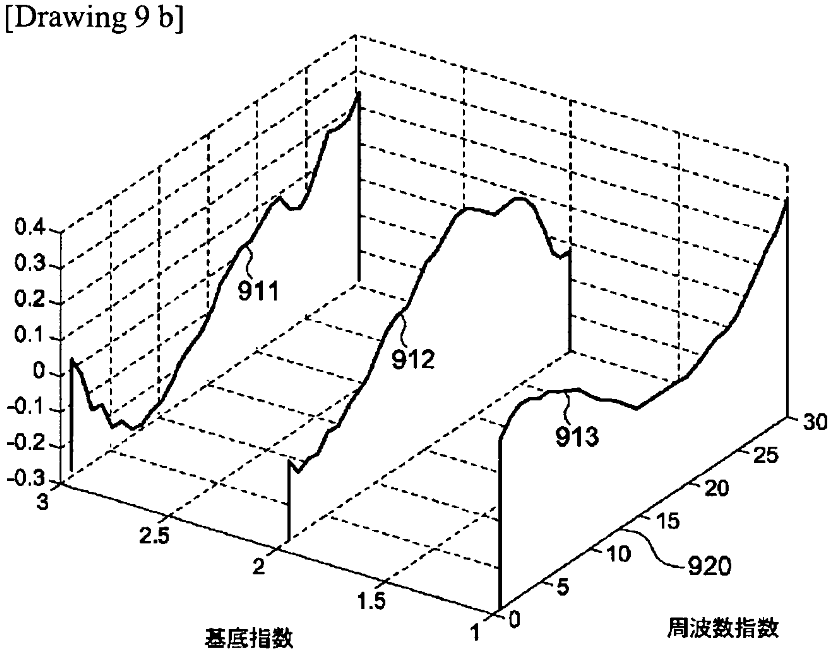
[Drawing 11 a]



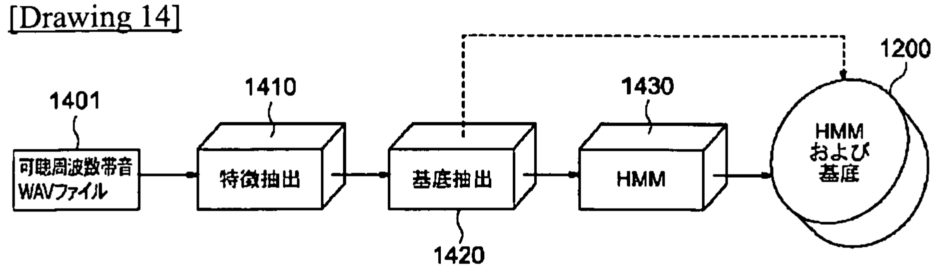
[Drawing 9 a]



[Drawing 9 b]

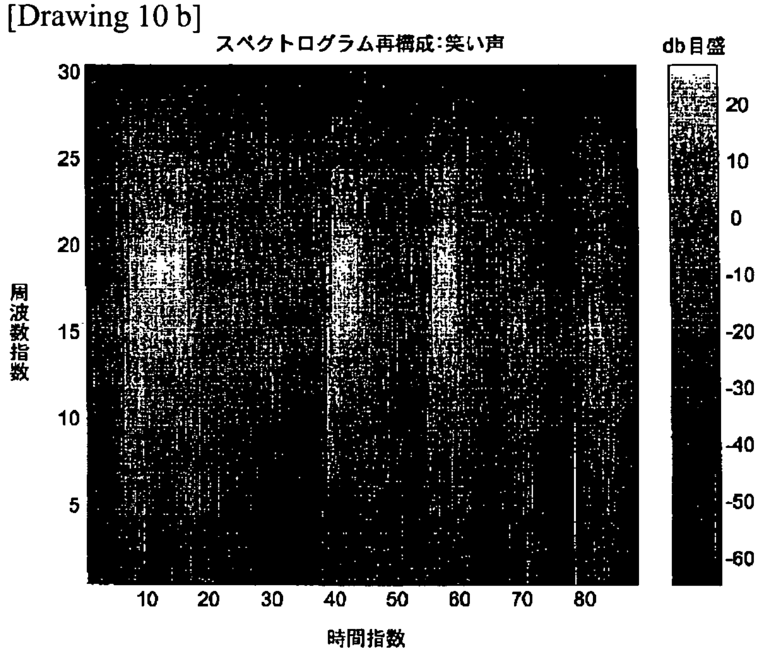


[Drawing 14]

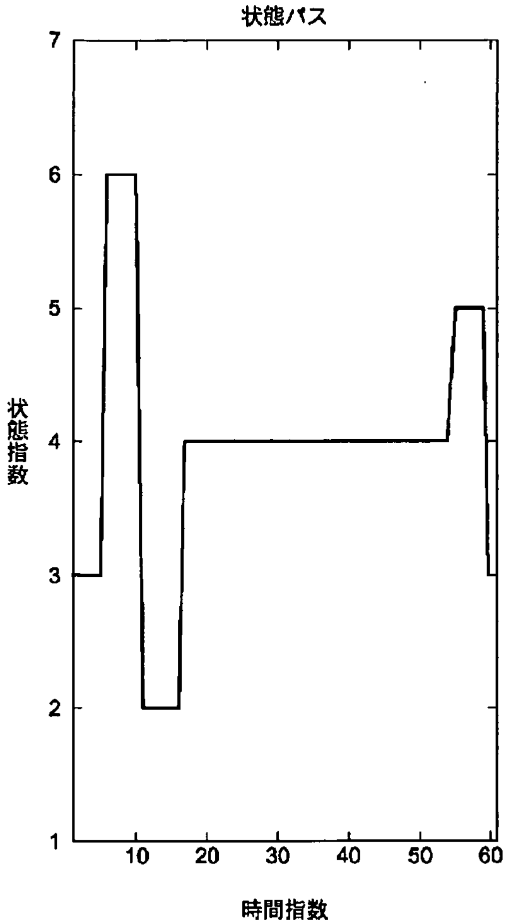


[Drawing 10 a]

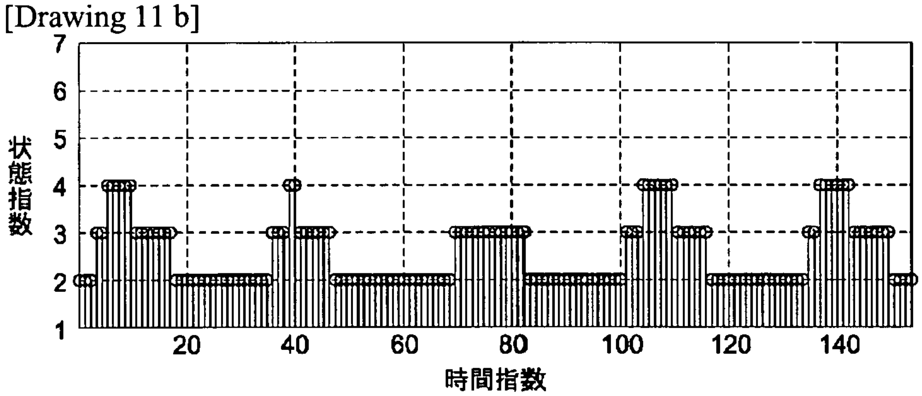
[Drawing 10 b]



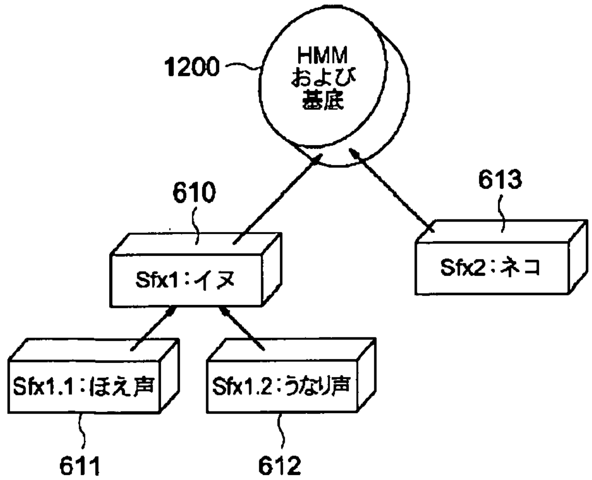
[Drawing 18 a]



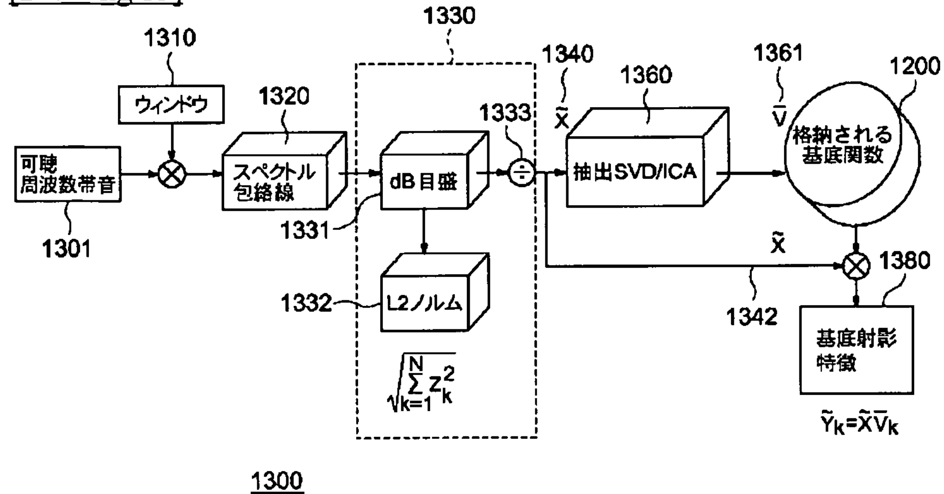
[Drawing 11 b]



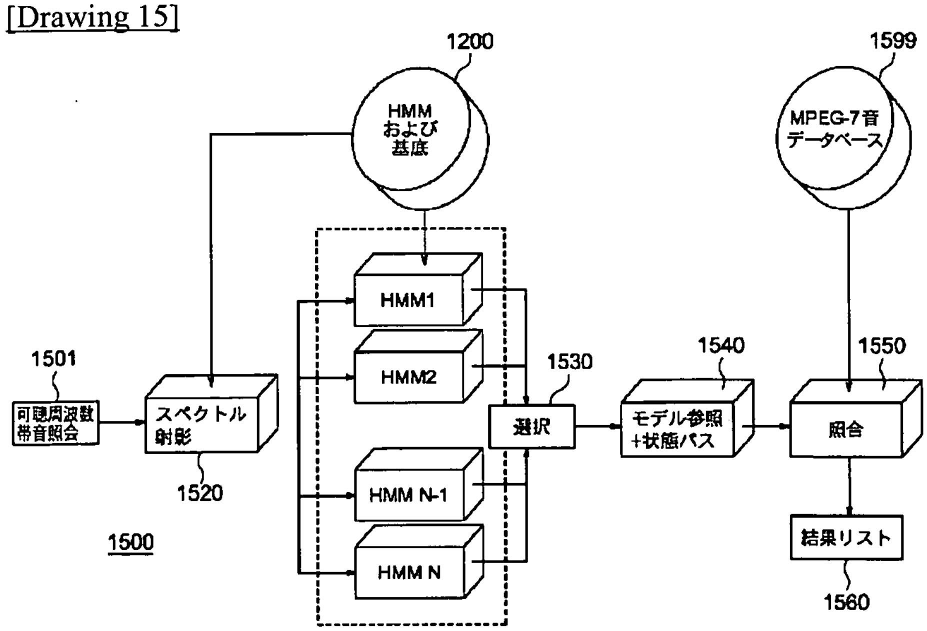
[Drawing 12]



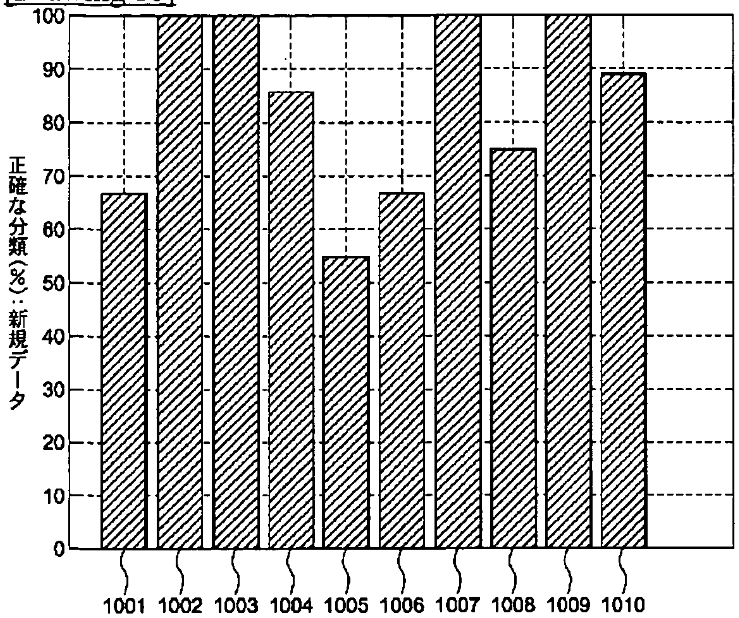
[Drawing 13]



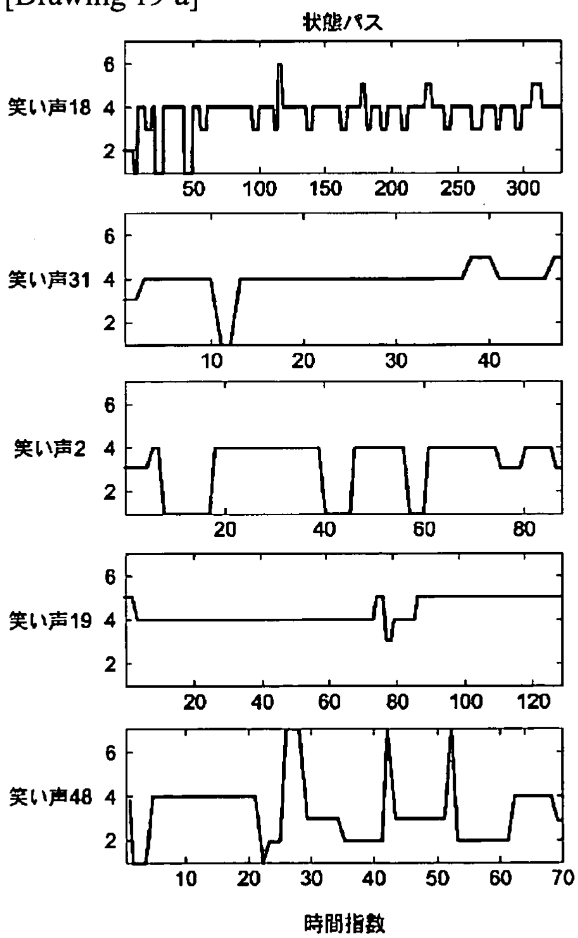
[Drawing 15]



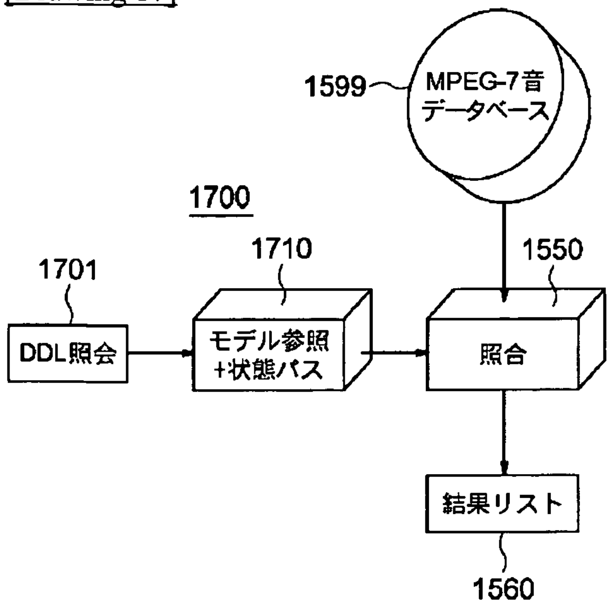
[Drawing 16]



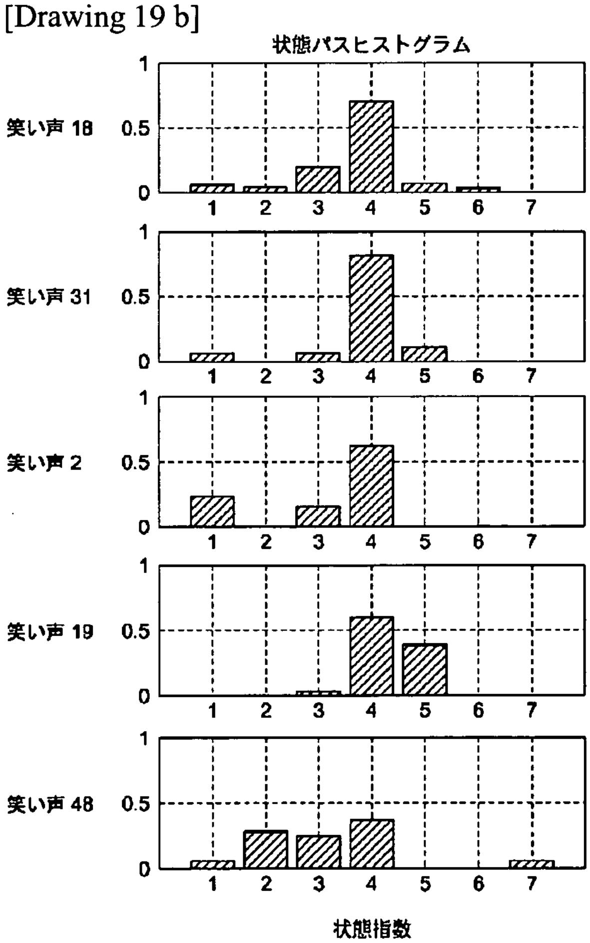
[Drawing 19 a]



[Drawing 17]



[Drawing 19 b]



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 June 2003 (12.06.2003)

PCT

(10) International Publication Number
WO 03/049432 A1

(51) International Patent Classification[7]: H04N 5/781

(21) International Application Number: PCT/US02/38395

(22) International Filing Date: 3 December 2002 (03.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/005,252 3 December 2001 (03.12.2001) US

(71) Applicant: SONY ELECTRONICS, INC. [US/US]; 1 Sony Drive, Park Ridge, NJ 07656 (US).

(72) Inventors: RISING, Hawley, K., III; 3294 Desertwood Lane, San Jose, CA 95132 (US). TABATABAI, Ali; 10495 SW 155th Avenue, Beaverton, OR 97007 (US).

(74) Agents: SALTER, James, H. et al.; Blakely, Sokoloff, Taylor & Zafman, 12400 Wilshire Boulevard, 7th Floor, Los Angeles, CA 90025-1026 (US).

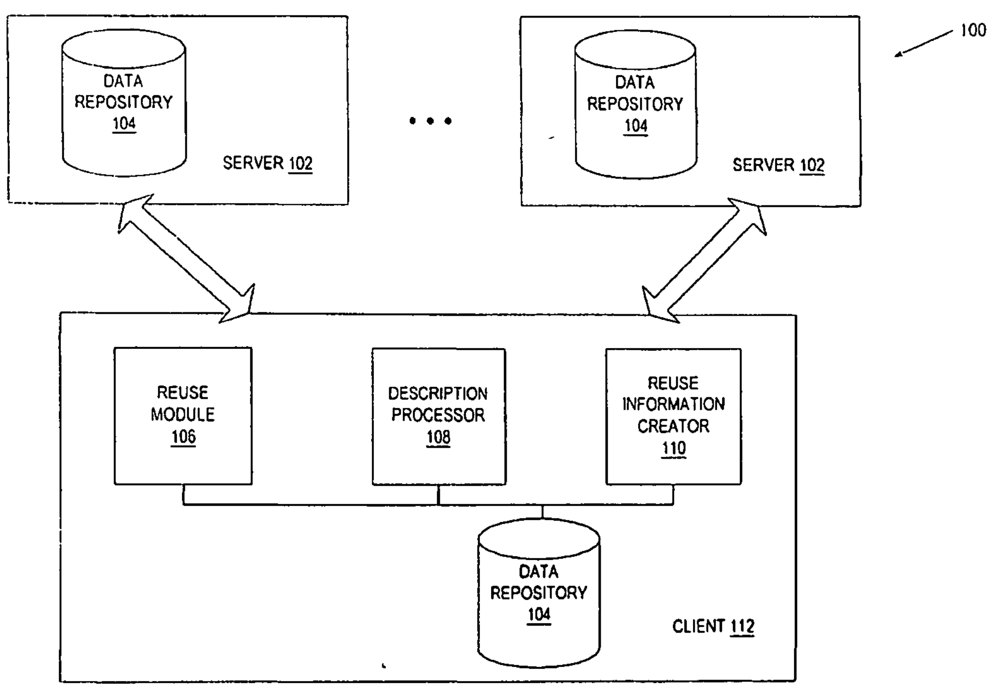
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

[Continued on next page]

(54) Title: LIMITED AUTHORIZATION



(57) Abstract: A method and apparatus for processing descriptions of audiovisual content (Fig. 4) are described. According to one embodiment, a description (104) of audiovisual content is created, and information pertaining to reuse of the description of audiovisual content is defined. Further, the description of audiovisual content and the reuse information are stored in a repository (104) of descriptive date to enable subsequent reuse of this description.

WO 03/049432 A1

DISTRIBUTED SEMANTIC DESCRIPTIONS OF AUDIOVISUAL CONTENT

FIELD OF THE INVENTION

The present invention relates to multimedia data processing and more specifically to multimedia data processing based upon semantic descriptions.

BACKGROUND OF THE INVENTION

The Motion Picture Expert Group (MPEG) develops standards concerning audiovisual content. One component of the MPEG standard scheme includes MPEG-7 standards which are directed to providing descriptions of audiovisual content that may be of interest to the user. Specifically, the MPEG-7 standards are developed to standardize information describing the audiovisual content. The MPEG-7 standards may be used in various areas, including storage and retrieval of audiovisual items from databases, broadcast media selection, tele-shopping, multimedia presentations, personalized news service on the Internet, etc.

According to MPEG-7 standards, descriptions of audiovisual content consist of descriptors and description schemes. Descriptors represent features of audiovisual content and define the syntax and the semantics of each feature representation. Description schemes (DS) specify the structure and semantics of the relationships between their components. These components may be both descriptors and description schemes. Conceptual aspects of a description scheme can be organized in a tree or in a graph. The graph structure is defined by a set of nodes that represent elements of a description scheme and a set of edges that specify the relationship between the nodes.

Descriptions (i.e., descriptors and DSs) of audiovisual content are divided into segment descriptions and semantic descriptions. Segment descriptions describe the audiovisual content from the viewpoint of its structure. That is, the descriptions are structured around segments which represent physical spatial, temporal or spatio-temporal components of the audiovisual content. Each

segment may be described by signal-based features (color, texture, shape, motion, audio features, etc.) and some elementary semantic information.

Semantic descriptions describe the audiovisual content from the conceptual viewpoints, i.e., the semantic descriptions describe the actual meaning of the audiovisual content rather then its structure. The segment descriptions and semantic descriptions are related by a set of links, which allows the audiovisual content to be described on the basis of both content structure and semantics together. The links relate different semantic concepts to the instances within the audiovisual content described by the segment descriptions.

Current semantic descriptions are limited in their descriptive capabilities because they describe specific semantic entities without identifying the relationships between these specific semantic entities and other related semantic entities. For instance, the current model of a semantic description includes multiple DSes for various semantic entities such as, for example, an event, an object, a state, an abstract concept, etc. An event DS describes a meaningful temporal localization. For example, an event DS may be associated with a concrete instance in the real world or the media (e.g., a wedding). An object DS describes semantically a specific object (e.g., a car depicted in an image). A state DS identifies semantic properties of the entity (e.g., of an object or event) at a given time, in a given spatial location, or in a given media location. A concept DS describes abstract elements that are not created by abstraction from concrete objects and events. Concepts such as freedom or mystery are typical examples of entities described by concept descriptions.

The above DSes describe specific entities. However, a description cannot be complete if it only describes an individual entity by itself. Most human description and communication is accomplished by bringing information together, information is seldom completely delineated in any exchange. Hints are present in speech that cause both parties to construct reasonably compatible or similar mental models, and the information discussed is discussed within such context. Accordingly, a description cannot accurately and completely describe the content unless it contains various additional information related to

this content. This additional information may include background information, context information, information identifying relationships between the content being described and other entities, etc.

In addition, no current mechanism exists for creating descriptions of metaphors or analogies. A traditional opinion is that semantic descriptions should only describe audiovisual material and, therefore, there is no need to create metaphorical descriptions. However, humans use metaphors and analogies all the time without realization of such use. Such metaphors and analogies as "feeling like a fish out of water," "getting close to the deadline," "flying like a bird," etc. are inherent in human communication. Thus, it would be undesirable to exclude descriptions of metaphors and analogies from a list of possible descriptions.

Further, current semantic descriptions are static. When the material described by an existing semantic description changes, the process of creating a description must be performed anew to produce a new semantic description describing the changed material.

Accordingly, a tool is required to create semantic descriptions that are capable of completely and accurately describe any semantic situation, audiovisual or otherwise. Such a tool should also be able to create descriptions that would dynamically reflect changes in the material being described.

SUMMARY OF THE INVENTION

A method and apparatus for processing descriptions of audiovisual content are described. According to one embodiment, a description of audiovisual content is created, and information pertaining to reuse of the description of audiovisual content is defined. Further, the description of audiovisual content and the corresponding reuse information are stored in a repository of descriptive data to enable subsequent reuse of this description.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation in the figures of the accompanying drawings in which like reference numerals refer to similar elements.

Figures 1 and 2 are prior art embodiments for creating mental spaces;

Figure 3 illustrates a hierarchy of various structural forms of semantic descriptions of audiovisual content;

Figure 4 is a block diagram of one embodiment of a system for processing descriptions of audiovisual content;

Figure 5 is a flow diagram of one embodiment for providing distributed descriptions of audiovisual content;

Figure 6 is a flow diagram of one embodiment for reusing descriptions of audiovisual content;

Figure 7 is a flow diagram of one embodiment for dynamic reuse of descriptions of audiovisual content;

Figure 8 illustrates an exemplary semantic mosaic; and

Figure 9 is a block diagram of one embodiment of a computer system.

DETAILED DESCRIPTION

A method and apparatus for processing descriptions of audiovisual content are described. In the following detailed description of the present invention, numerous specific details are set forth in order to provide a thorough understanding of the present invention. However, it will be apparent to one skilled in the art that the present invention may be practiced without these specific details. In some instances, well-known structures and devices are shown in block diagram form, rather than in detail, in order to avoid obscuring the present invention.

Some portions of the detailed descriptions that follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most

effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like.

It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussion, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

The present invention also relates to apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general purpose computer selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, and magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus.

The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general purpose systems may be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these systems will appear from the description below. In addition, the present invention is not described with reference to any particular programming language. It will be appreciated that a variety of programming languages may be used to implement the teachings of the invention as described herein.

The present invention relates to various aspects of creating descriptions of audiovisual content. As described above, current descriptive tools lack the ability to produce descriptions that can describe audiovisual content in a complete and accurate manner. This limitation is caused by the entity-specific nature of current semantic descriptions. That is, each semantic description describes a specific entity independently from contextual and background information that relates to this specific entity. However, this approach contradicts the rules governing perception and interpretation of speech which is the prototype for semantic descriptions. Some of these rules are based on the use of mental space theory that is well known in the art.

Mental spaces provide context for communication by importing a lot of information not included in the speech, thereby providing a mechanism for interpreting semantic content in language. This information is imported using maps. These maps function by using (i.e., "recruiting") frames which represent predefined constructs for interpretation, projecting structure from one mental space to another, and integrating or abstracting imported material from more than one other mental space. Accordingly, each mental space may represent an extended description containing entities, relationships, and frames. Several mental spaces may be active at once, in order to properly define all the entities in the description. These mental spaces enter into relationships with each other. Because the mental spaces borrow structure and entities from each other, mappings are necessary between these mental spaces. The whole composite

forms a backdrop to the expressed description and completes the process of attaching semantic meaning to the entities involved.

Figures 1 and 2 are prior art embodiments for creating mental spaces. Referring to **Figure 1**, a new mental space 50 is created by recruiting some of frames 10 and borrowing structures from existing mental spaces 20 and 30. Referring to **Figure 2**, a new mental space 70 is created by blending or integrating two existing mental spaces 62 and 64. A generic space 66 may then be created by abstracting from all three mental spaces: new mental space 70 and existing mental spaces 64 and 62. A generic space 66 includes structures that are common to the mental spaces 62, 64 and 70.

The present invention uses the mental space model to enable creation of semantic descriptions that are capable of completely describing any semantic situation. In one embodiment, the tool for creating complete semantic descriptions is based on a number of description categories. These categories may include objects, events, states, episodes, frames, descriptive structures, and mental spaces. The term "object" as used herein refers to a description of a real object, or of a composite or abstraction of several real objects. Objects contain states. Each state is a collection of attributes that are attached to objects and relationships. By extension, states may be attribute collections of mental spaces. Objects may have subobjects and object states may have substates. A change in state is an event. As such, an event may be a change in any of the constituents of a description of an object or relationship (including what represents the mental spaces). Because states may have substates, events may have subevents.

States may also be attached to episodes, and an event may be a change in any of the constituents of a description of an episode. An episode is a semantically significant time span. Episodes may coincide with the behavior of objects, with the occurrence of events, with changes in relationships, or changes in the mental spaces used to provide context to the objects, events, and relationships. If semantically significant time spans are properly contained in an episode, they are subepisodes.

A frame is a presumed or predefined set of rules for interpreting or describing a set of semantic objects. As such, frames may be prototypical descriptions themselves, or they may be sets of rules, definitions, and descriptive structures. Descriptive structures are abstractions of objects, episodes, states, and relationships. A mental space is a collection of objects, relationships, and frames, together with mappings that embed descriptive structures from descriptions or from other mental spaces.

A complete description of semantic content may consists of any combination of descriptions of the above categories. In addition, the complete description should include descriptions of relationships between semantic entities that are included in the complete description of semantic content. A relationship between the entities is either a relation or a mapping. Because relations can be expressed as compositions of mappings, the term "mapping" can also be used to identify a relation. Relationships may be of various types such as, for example, inclusion, containment, similarity, example of, relative position, etc.

The relationships between objects form structure. Further, the mapping of objects, states, and events into an episode is structure. The mappings that make up the underlying mental spaces are structures. States may be represented as maps from the entities described by the states to spaces of attribute values. Even objects can be described as structure: objects are in one-to-one correspondence with the mappings from any point set to the objects themselves, or any mappings from the objects themselves to a one point set). Thus, structure is an inherent part of a semantic description construct.

Structure may take various forms including morphisms, graphs, categories, functors, natural transformations, etc. Morphisms are arrows between objects in a category. A category consists of two sets, a set of objects, and a set of morphisms, which obey the following two rules:

1) For each object, there must be a morphism to itself called the "identity" morphism;

2) If *f* is a morphism from A to B, and g is a morphism from B to C, then

there must be a morphism (usually written as $(g \circ f)$) from A to C that is equivalent to first determining f and then determining g.

It is possible to define mappings between categories. Mappings between categories must take objects to objects and morphisms to morphisms. Mappings between categories also need to take the source and target of any morphism to the source and target of its image under the mapping (this is a rule that defines morphisms for the category of graphs). Mappings between categories need to satisfy two constraints, called categorical constraints:

1) They must take the identity maps to identity maps; and

2) They must preserve compositions, i.e., if F takes A to X, B to Y, C to Z, and takes f to h and g to p, then F must take $(g \circ f)$ to $(p \circ h)$.

Any map that obeys the above constraints is called "categorical".

A categorical map between two categories is called a functor. A functor maps between categories, e.g., F maps category C to category D. It is possible to see this with C and D being like objects, and F being like an arrow (morphism). If G maps category H to category J, then we can make a new map that takes C to H, D to J and F to H. If this new map obeys categorical constraints, then it is called a Natural Transformation.

Figure 3 illustrates a hierarchy of various structural forms of semantic descriptions of audiovisual content. Morphism 302 is a map between two objects 304. Each category 310 consists of a set of objects (including, for example, objects 304) and a set of morphisms (including, for example, morphism 302). Functor 306 is a map between categories 310. Natural transformation 308 is a map between functors. There is no need to make a map between natural transformations because the hierarchy can be continued using "functor categories").

Thus, a complete description of audiovisual content may include descriptions of various semantic entities (e.g., objects, events, states, episodes, frames, descriptive structures, and mental spaces), together with descriptions expressing the structure of the complete description. Although this approach provides semantic descriptions that are capable of describing any semantic

description in a complete and accurate manner, it may add a significant degree of complexity to the resulting semantic descriptions. One embodiment of the present invention addresses this complexity by distributing existing descriptions of audiovisual content. In this embodiment, existing descriptions can be archived and then reused to create new descriptions, as will be described in greater detail below.

Figure 4 is a block diagram of one embodiment of a system 100 for processing descriptions of audiovisual content. System 100 consists of one or more server computers 112 coupled to one or more client computers such as client 112. Client 112 may communicate with server 102 via any wire or wireless communication link including, for example, a public network such as Internet, a local network such as Ethernet, Intranet and local area network (LAN), or a combination of networks. Each of client 112 and server 102 may be any type of computing device such as, for example, a desktop computer, a workstation, a laptop, a mainframe, etc.

In one embodiment, server 102 contains data repository 104 which stores various descriptions of audiovisual content. In one embodiment, data repository 104 contains only semantic descriptions of audiovisual content, i.e., descriptions that describe the actual meaning of the audiovisual content. Alternatively, data repository 104 stores descriptions of other types (e.g., segment descriptions), in addition to semantic descriptions.. Descriptions are stored independently from the audiovisual content that they describe. In one embodiment, each description is stored with associated reuse information which indicates how this description can be reused to create other descriptions of audiovisual content. The functionality of the reuse information will be described in greater detail below.

Client 112 includes a tool for creating new descriptions by reusing existing descriptions of audiovisual content. In one embodiment, this tool includes a reuse module 106, a description processor 108, and a reuse information creator 110. In one embodiment, client 112 also includes a data repository 114 to store descriptions of audiovisual content locally.

Reuse module 106 is responsible for finding existing descriptive data that can be reused to create a new description of audiovisual content. In one embodiment, this descriptive data resides in data repository 104 of one or more servers 102. Alternatively, some or all of this descriptive data may reside locally in data repository 114. The existing descriptive data may include portions or entire descriptions of audiovisual data. As described above, each description is stored with associated reuse information. The reuse module 106 is responsible for analyzing this reuse information to determine what type of reuse is allowable for this particular description.

Description processor 108 is responsible for creating new descriptions of audiovisual content using the existing descriptive data and the associated reuse information. Reuse information creator 119 is responsible for defining reuse information for the newly created description of audiovisual content. In one embodiment, the new description is stored locally in data repository 114. Alternatively, the new description is transferred to server 102 for storing in data repository 104. In either embodiment, the new description is stored with associated reuse information to enable subsequent reuse of this description.

Figure 5 is a flow diagram of one embodiment for providing distributed descriptions of audiovisual content. At processing block 504, a new description of audiovisual content is created. In one embodiment, the new description is created by reusing one or more existing descriptions as will be described in greater detail below in conjunction with **Figure 6**. Alternatively, a new description is created by abstracting from a plurality of existing descriptions, i.e., by extracting common attributes from the existing descriptions. In one embodiment, the new description is a descriptor. Alternatively, the new description is a description scheme (DS). As described above, descriptors represent features of audiovisual content and define the syntax and the semantics of each feature representation. DSes specify the structure and semantics of the relationships between their components. These components may be both descriptors and description schemes. In one embodiment, the new description is a semantic description. A semantic description may describe such

semantic entities as events, objects, states, relationships, episodes, descriptive structures, mental spaces, or any combination of the above semantic entities.

At processing block 506, information pertaining to subsequent reuse of the created description is defined. This information indicates what type of reuse is allowable for this description. For example, the reuse information may indicate whether this description can be embedded in another description without changing the intended meaning of this description or whether this description can be subdivided into components which maintain their meaning when extracted for reuse. The reuse information may also indicate whether the description can be transformed to enable the reuse of this description. For example, the reuse information may specify whether a description of an eye can be mirrored to produce a description of the other eye. Further, the reuse information may indicate whether the description can maintain its transitive capability when this description is reused. For example, the reuse information may specify whether the description will function as a subset if this description is embedded into a larger description.

At processing block 508, the description and associated reuse information are stored in a repository of descriptive data to enable subsequent reuse of this description. The reuse information may be stored as a set of flags associated with various reuse types, as a number specifying a combination of reuse types allowable for the description, or in any other form. In one embodiment, the description is stored on a network server and may be accessed by a plurality of client computers over a network (e.g., Internet or a local network). Alternatively, the description may be stored locally on a client computer and may be accessed by the users of the client computer. In either embodiment, the description can subsequently be reused to create new descriptions based on the reuse information associated with this description.

Figure 6 is a flow diagram of one embodiment for reusing descriptions of audiovisual content. At processing block 604, existing descriptive data that should be included in a new description is found. In one embodiment, the existing descriptive data includes one or more descriptions of audiovisual

content (or portions of descriptions) that are selected from a plurality of descriptions stored on a network server(s). For example, a description provider may create a plurality of descriptions that may potentially have a widespread use and publish them on a web site for future reuse. In another example, descriptions published on a web site may be abstractions (or templates) created by extracting common features from various existing descriptions. In this example, such description may be stored with an indicator specifying that this description is an abstraction. In another embodiment, the existing descriptive data or its portion is selected from a local repository of descriptive data.

At processing block 606, reuse information associated with the selected descriptive data is analyzed to determine how the selected descriptive data can be reused. As described above, the reuse information may indicated whether the selective descriptive data can be subsumed, subdivided or transformed, or whether the selected descriptive data is transitive.

At processing block 608, a new description is created using the selected descriptive data and associated reuse information. In one embodiment, the new description includes a reference to the selected descriptive data, rather than the date itself, thereby avoiding the creation of a large and complex description. Since the descriptive data may consist of multiple descriptions (or their portions), the description may include references to multiple descriptions. For instance, a new DS may include references to such DSes as, for example, object DSes, event DSes, state DSes, relationship DSes, episode DSes, descriptive structure DSes, and mental space DSes. Depending on the form of reuse, a mapping of each existing description into a new description is needed. In one embodiment, such mapping is defined each time it is needed to create a new description. Alternatively, an archived version of the mapping is referred to in a new description.

In one embodiment, the new description is created by converting the existing descriptive data into a part of a description and mapping this partial description into a new description. For instance, under current MPEG-7 standards, a complete semantic description may include multiple object DSes,

event DSes, and concept DSes. A concept DS, which is intended to allow encapsulation of a complex abstraction, may again contain object DSes, event DSes, and concept DSes. Since a concept DS can be included in descriptions of objects and events, creating a new description of an object or event requires converting an existing concept DS into a part of the new description and mapping this concept DS into the new description.

In another embodiment, a new description is created by accessing a portion of an existing description and mapping this partial description into the new description, thereby enabling the reuse of a portion of an existing description, rather than the entire existing description. For instance, an object DS contained within an existing concept DS may be accessed and mapped into a new description of audiovisual material. In one embodiment, a partial description is extracted from an existing description, converted into a standalone description, and then embedded into a new description.

In yet another embodiment, a new description is created by selecting various existing descriptions (or their portions), and combining them by using combination rules from a dictionary of rules for combining descriptions. The existing descriptions are mapped into the dictionary entries, and the rules are executed to create a new description. Then, the corresponding objects are identified with parts of the new description. The rules and descriptions can be located on the local machine, in a single data repository, or in several data repositories, and may be executed by the description processor. The data repositories may have rules for forming descriptions as well as existing descriptions to use, and these are organized in dictionaries.

In one embodiment, existing descriptions or portions of existing descriptions are mapped into new descriptions using any known in the art mechanisms that are capable of performing graph operations between various descriptions of audiovisual data. Alternatively, object oriented inheritance mechanisms may be used for this purpose. For example, private inheritance allows the inheritance of attributes and methods without the acquisition of a

data type relationship. Accordingly, private inheritance can be used, for example, to map a portion of an existing description into a new description. Public inheritance provides a mechanism for generating categorical structure. Thus, public inheritance can be used, for example, to map an existing description, which is converted into a part of a new description, to the new description. In addition, both private inheritance and public inheritance can be used to map existing descriptions to new descriptions. For example, both types of inheritance may be used to map existing descriptions into abstractions and then to further map the abstractions into a new description which combines these abstractions.

In one embodiment, multiple reuse of descriptions enables de facto standardization (as opposed to pure standardization) of the descriptions by category. That is, as opposed to the pure standardization imposed as the outset in MPEG-7 that cannot possibly know the categories arising in application after the standard is adopted, the standard may be created naturally by identifying, through multiple reuse, those description categories that have the most application and use.

According to one embodiment of the present invention, the creation of descriptions of audiovisual content is performed at the same time as the creation of the audiovisual content by dynamically reusing existing descriptions. For example, when describing a news program dedicated to a national disaster concurrently with the continuing coverage of the national disaster, the descriptions of the news program are dynamically updated to create new descriptions of the evolving content.

Figure 7 is a flow diagram of one embodiment for dynamic reuse of descriptions of audiovisual content. At processing block 704, a first description of audiovisual content is created. At processing block 706, reuse information associated with the first description is defined as described in more detail above. In one embodiment, the first description and the associated reuse information is then stored in a local data repository.

At processing block 708, the first description is reused to create a second description of modified audiovisual content based on the reuse information. The second description is created concurrently with the creation of the modified audiovisual content. In one embodiment, the second description is created by updating parameter values of the first description. In another embodiment, the second description is created by combining the first description with other new or existing descriptive data. In yet another embodiment, the second description is created by reusing some portions of the first description and discarding the other portions of the first description that are no longer appropriate. For example, during the description of an online episode, the relationships between its objects may vary, as well as the structures needed to describe them. Then, the current description may need to be modified by updating its parameter values and by bringing in new descriptions or partial descriptions to describe emergent behavior, discarding portions of the current description that are no longer needed.

In one embodiment, dynamic reuse is performed using object-oriented modeling such as system object model (SOM) of IBM™. SOM, which is an architecture that allows binary objects to be shared by different applications, enables dynamic changes of descriptions, relationships, and attributes of a structure while it evolves.

One embodiment of the present invention utilizes a semantic mosaic to create new descriptions of audiovisual content. A semantic mosaic is a collection of various descriptions that are blended together using interrelations between neighboring descriptions. **Figure** 8 illustrates an exemplary semantic mosaic 800. Mosaic 800 is composed of multiple semantic descriptions of audiovisual content, including descriptions 1-18. When mosaic 800 is created, descriptions that relate to each other are blended. For example, descriptions 1 and 2 have a point at which they carry the same information. This point is used to blend descriptions 1 and 2 together. Description 2 may also have another common point with description 3. This other point may be used to blend description 3 with description 2. A third common point may be used to blend descriptions 2

and 5 together, etc. As a result, description 2 is blended with descriptions 1, 3, 4 and 5 that all relate to description 2 but may not have any interrelations between each other. Thus, semantic mosaic 800 presents a description which does not describe any semantic material in particular but includes local pieces that represent descriptions of various semantic content. Each local piece may combine several descriptions to describe a certain semantic entity. Depending on the context, the number of the descriptions included in a particular local piece may vary. For instance, in one context, the combination of descriptions 5, 10 and 11 may provide a complete description of audiovisual content. In another context, the combination of descriptions 5, 9, 10, 12 and 13 may be needed to provide a complete description of audiovisual content. When a new description is created, an appropriate local piece may be reused to create the new description. The descriptions contained in each local piece have previously defined relationships. Thus, new descriptions may be created by merely extracting appropriate local pieces from the semantic mosaic. Alternatively, the local pieces may be combined with other descriptive data to form new descriptions.

Figure 9 is a block diagram of one embodiment of a computer system 900 within which a set of instructions, for causing the machine to perform any one of the methodologies discussed above, may be executed. In alternative embodiments, the machine may comprise a network router, a network switch, a network bridge, Personal Digital Assistant (PDA), a cellular telephone, a web appliance or any machine capable of executing a sequence of instructions that specify actions to be taken by that machine.

The computer system 900 includes a processor 902, a main memory 904 and a static memory 906, which communicate with each other via a bus 908. The computer system 900 may further include a video display unit 910 (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)). The computer system 900 also includes an alpha-numeric input device 912 (e.g., a keyboard), a cursor control device 914 (e.g., a mouse), a disk drive unit 916, a signal generation device 920 (e.g., a speaker) and a network interface device 922.

The disk drive unit 916 includes a computer-readable medium 924 on which is stored a set of instructions (i.e., software) 926 embodying any one, or all, of the methodologies described above. The software 926 is also shown to reside, completely or at least partially, within the main memory 904 and/or within the processor 902. The software 926 may further be transmitted or received via the network interface device 922. For the purposes of this specification, the term " computer-readable medium" shall be taken to include any medium that is capable of storing or encoding a sequence of instructions for execution by the computer and that cause the computer to perform any one of the methodologies of the present invention. The term "computer-readable medium" shall accordingly be taken to include, but not be limited to, solid-state memories, optical and magnetic disks, and carrier wave signals.

Thus, a method and apparatus for processing descriptions of audiovisual content have been described. Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the invention. Accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense.

CLAIMS

What is claimed is:

1. A method for processing descriptions of audiovisual content, the method comprising:

   creating a first description of audiovisual content (504);
   defining information pertaining to reuse of the first description (506); and
   storing the first description and the information pertaining to reuse of the first description (508) in a repository (104) of descriptive data to enable subsequent reuse of the first description.

2. The method of claim 1 wherein the first description is a semantic description.

3. The method of claim 1 wherein the first description is a description scheme.

4. The method of claim 1 wherein the information pertaining to reuse of the first description indicates whether the first description can be embedded into a second description of audiovisual content without changing an intended meaning of the first description.

5. The method of claim 1 wherein the information pertaining to reuse of the first description indicates whether the first description can be divided into a plurality of partial descriptions, each of the plurality of partial descriptions being suitable for subsequent reuse.

6. The method of claim 1 wherein the information pertaining to reuse of the first description indicates whether the first description can be transformed when reused to create a second description of audiovisual content.

7. The method of claim 1 wherein the information pertaining to reuse of the first description indicates whether the first description can maintain transitive capability if the first description is reused to create a second description of audiovisual content.

8. The method of claim 1 further comprising:

reusing a plurality of descriptions (708) stored in one or more repositories (104) of descriptive data a number of times to provide de facto standardization of the plurality of descriptions by category.

9. A method for reusing descriptions of audiovisual content, the method comprising:

finding existing descriptive data (604) that should be included in a new description of audiovisual content;

analyzing reuse information associated with the descriptive data (606); and

creating the new description (608) using the existing descriptive data and the associated reuse information.

10. The method of claim 9 wherein the new description is a semantic description.

11. The method of claim 9 wherein the new description is a description scheme.

12. The method of claim 9 wherein the descriptive data is at least a portion of one or more existing descriptions of audiovisual content.

13. The method of claim 9 further comprising:

retrieving the descriptive data from one or more repositories (104) of descriptive data.

14. The method of claim 9 wherein creating the new description further comprises:

converting the existing descriptive data into a partial description; and

mapping the partial description to the new description.

15. The method of claim 9 wherein creating the new description further comprises: accessing a portion of the existing descriptive data in a repository (104) of descriptive data; and

mapping the portion of the existing descriptive data to the new description.

16. The method of claim 9 wherein creating the new description further comprises: performing dictionary mapping of objects in the existing descriptive data to corresponding objects in the new description.

17. The method of claim 9 wherein creating the new description further comprises:

including a reference to the existing descriptive data into the new description.

18. The method of claim 9 wherein the new description is created using a mechanism for performing graph operations.

19. The method of claim 9 wherein the new description is created using an object oriented inheritance mechanism.

20. The method of claim 9 wherein creating the new description further comprises:

extracting the existing descriptive data from a semantic mosaic that integrates a plurality of related descriptions.

21. A method for dynamically reusing descriptions of audiovisual content, the method comprising:

creating a first description of audiovisual content (704);

defining reuse information associated with the first description (706); and

reusing the first description (708) to create a second description of modified audiovisual content based on the reuse information, the reuse being performed concurrently with creation of the modified audiovisual content.

22. A system for processing descriptions of audiovisual content, the system comprising:

means for creating a first description of audiovisual content (108);

means for defining information pertaining to reuse of the first description (110); and

means for storing the first description and the information pertaining to reuse of the first description in a repository (104) of descriptive data to enable subsequent reuse of the first description.

23. An apparatus comprising:

a description processor (108) to create a first description of audiovisual content;

a reuse information creator (110) to define information pertaining to reuse of the first description; and

a repository (104) of descriptive data to store the first description and the information pertaining to reuse of the first description to enable subsequent reuse of the first description.

24. The apparatus of claim 23 wherein the first description is a semantic description.

25. The apparatus of claim 23 wherein the first description is a description scheme.

26. The apparatus of claim 23 wherein the information pertaining to reuse of the first description indicates whether the first description can be embedded into a second description of audiovisual content without changing an intended meaning of the first description.

27. The apparatus of claim 23 wherein the information pertaining to reuse of the first description indicates whether the first description can be divided into a plurality of partial descriptions, each of the plurality of partial descriptions being suitable for subsequent reuse.

28. The apparatus of claim 23 wherein the information pertaining to reuse of the first description indicates whether the first description can be transformed when reused to create a second description of audiovisual content.

29. The apparatus of claim 23 wherein the information pertaining to reuse of the first description indicates whether the first description can maintain transitive capability if the first description is reused to create a second description of audiovisual content.

30. A system for reusing descriptions of audiovisual content, the system comprising:

means for finding existing descriptive data (106) that should be included in a new description of audiovisual content;

means for analyzing reuse information associated with the descriptive data (106); and

means for creating the new description (108) using the existing descriptive data and the associated reuse information.

31. An apparatus comprising:

a reuse module (106) to find existing descriptive data that should be included in a new description of audiovisual content and to analyze reuse information associated with the descriptive data; and

a description processor (108) to create the new description using the existing descriptive data and the associated reuse information.

32. The apparatus of claim 31 wherein the new description is a semantic description.

33. The apparatus of claim 31 wherein the new description is a description scheme.

34. The apparatus of claim 31 wherein the descriptive data is at least a portion of one or more existing descriptions of audiovisual content.

35. The apparatus of claim 31 wherein the new description is created using a mechanism for performing graph operations.

36. The apparatus of claim 31 wherein the new description is created using an object oriented inheritance mechanism.

37. A system for dynamically reusing descriptions of audiovisual content, the method comprising:

means for creating a first description of audiovisual content (108);

means for defining reuse information associated with the first description (110); and

means for reusing the first description (108) to create a second description of modified audiovisual content based on the reuse information, the reuse being performed concurrently with creation of the modified audiovisual content.

38. An apparatus comprising:

a description processor (108) to create a first description of audiovisual content; and

a reuse information creator (110) to define reuse information associated with the first description, the description processor (108) to reuse the first description to create a second description of modified audiovisual content based on the reuse information, the reuse being performed concurrently with creation of the modified audiovisual content.

39. A computer readable medium that provides instructions, which when executed on a processor, cause said processor to perform operations comprising:

creating a first description of audiovisual content (504);

defining information pertaining to reuse of the first description (506); and

storing the first description and the information pertaining to reuse of the first description (508) in a repository (104) of descriptive data to enable subsequent reuse of the first description.

40. A computer readable medium that provides instructions, which when executed on a processor, cause said processor to perform operations comprising:

finding existing descriptive data (604) that should be included in a new description of audiovisual content;

analyzing reuse information associated with the descriptive data (606); and

creating the new description (608) using the existing descriptive data and the associated reuse information.

41. A computer readable medium that provides instructions, which when executed on a processor, cause said processor to perform operations comprising:

creating a first description of audiovisual content (704);

defining reuse information associated with the first description (706); and

reusing the first description (708) to create a second description of modified audiovisual content based on the reuse information, the reuse being

performed concurrently with creation of the modified audiovisual content.

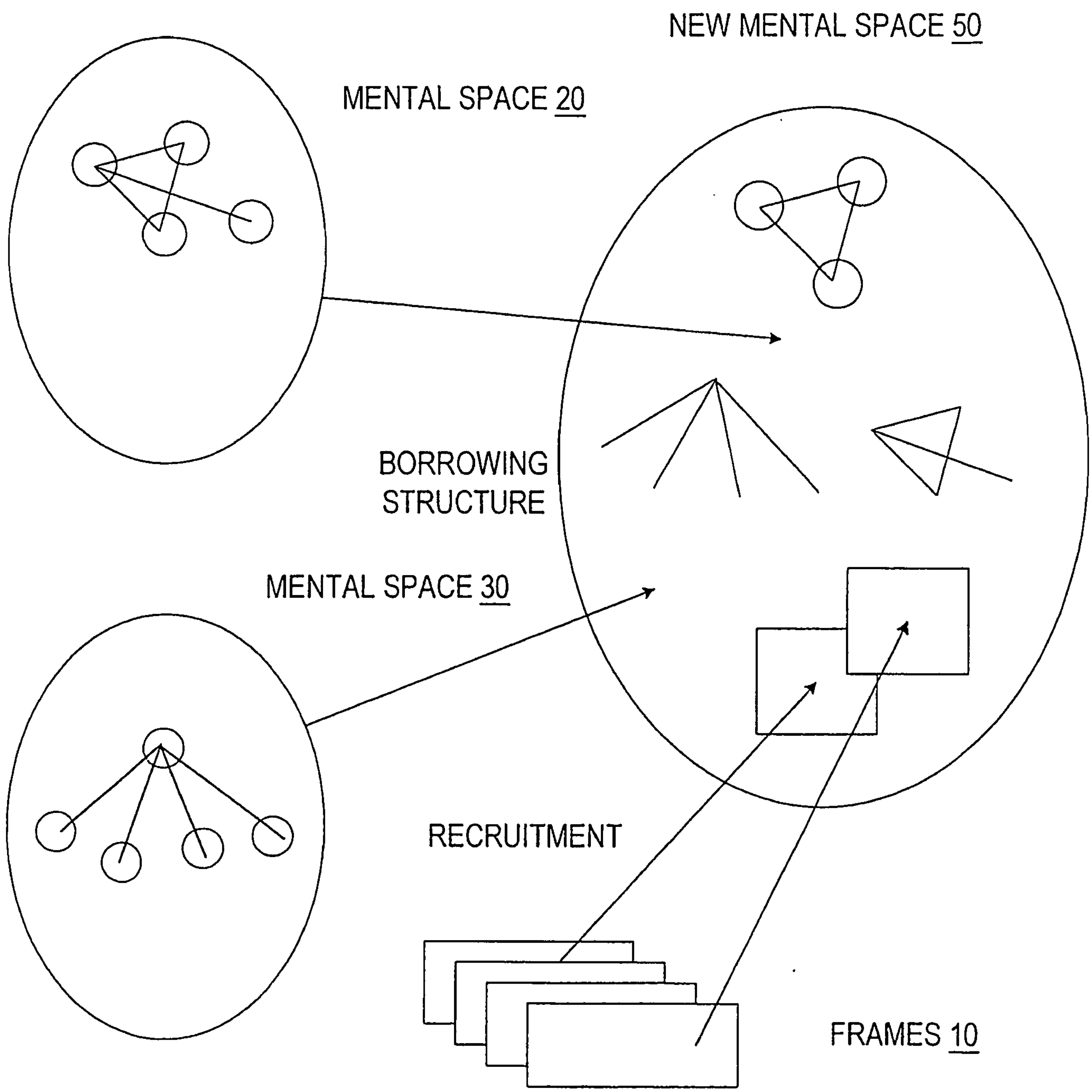


FIG. 1
PRIOR ART

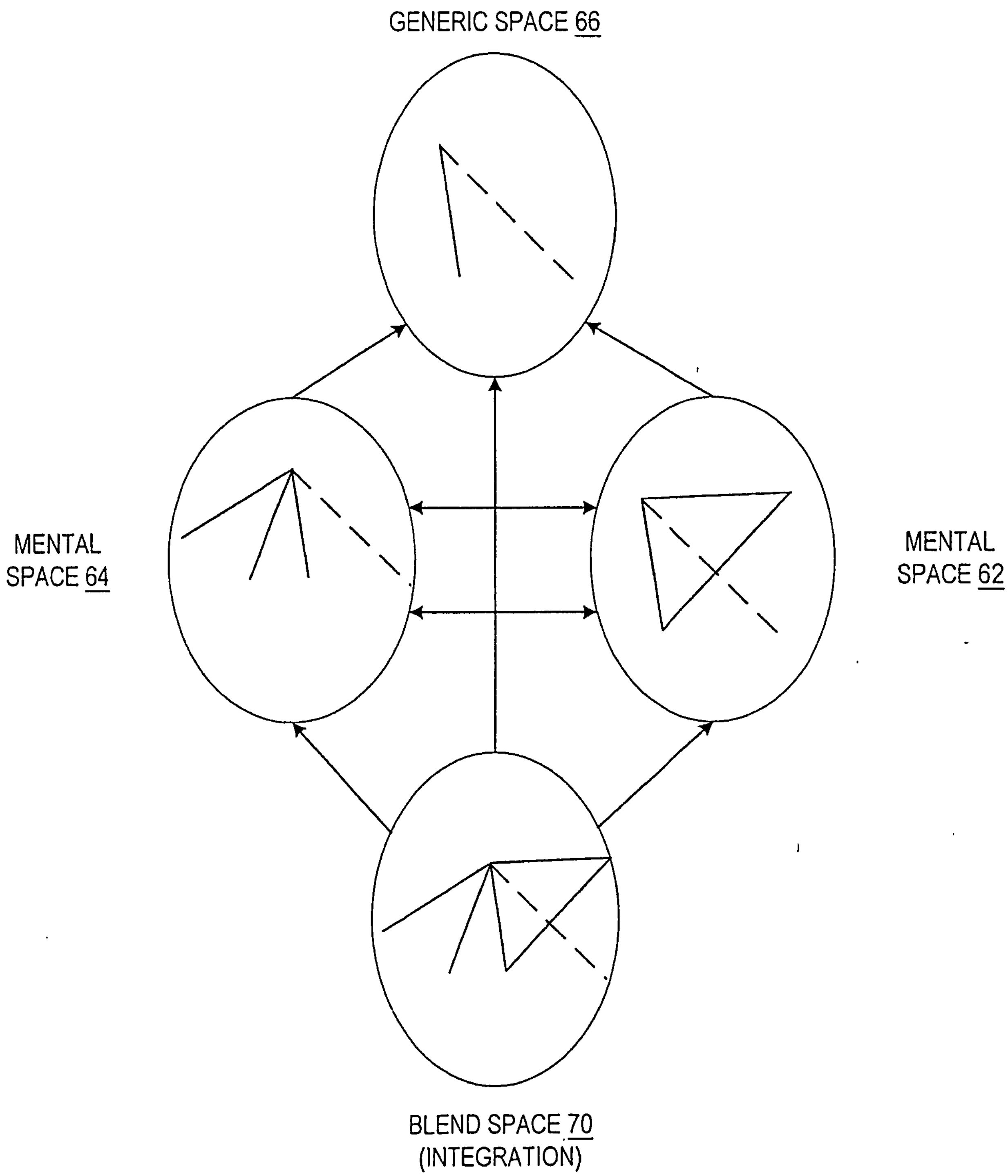


FIG. 2
PRIOR ART

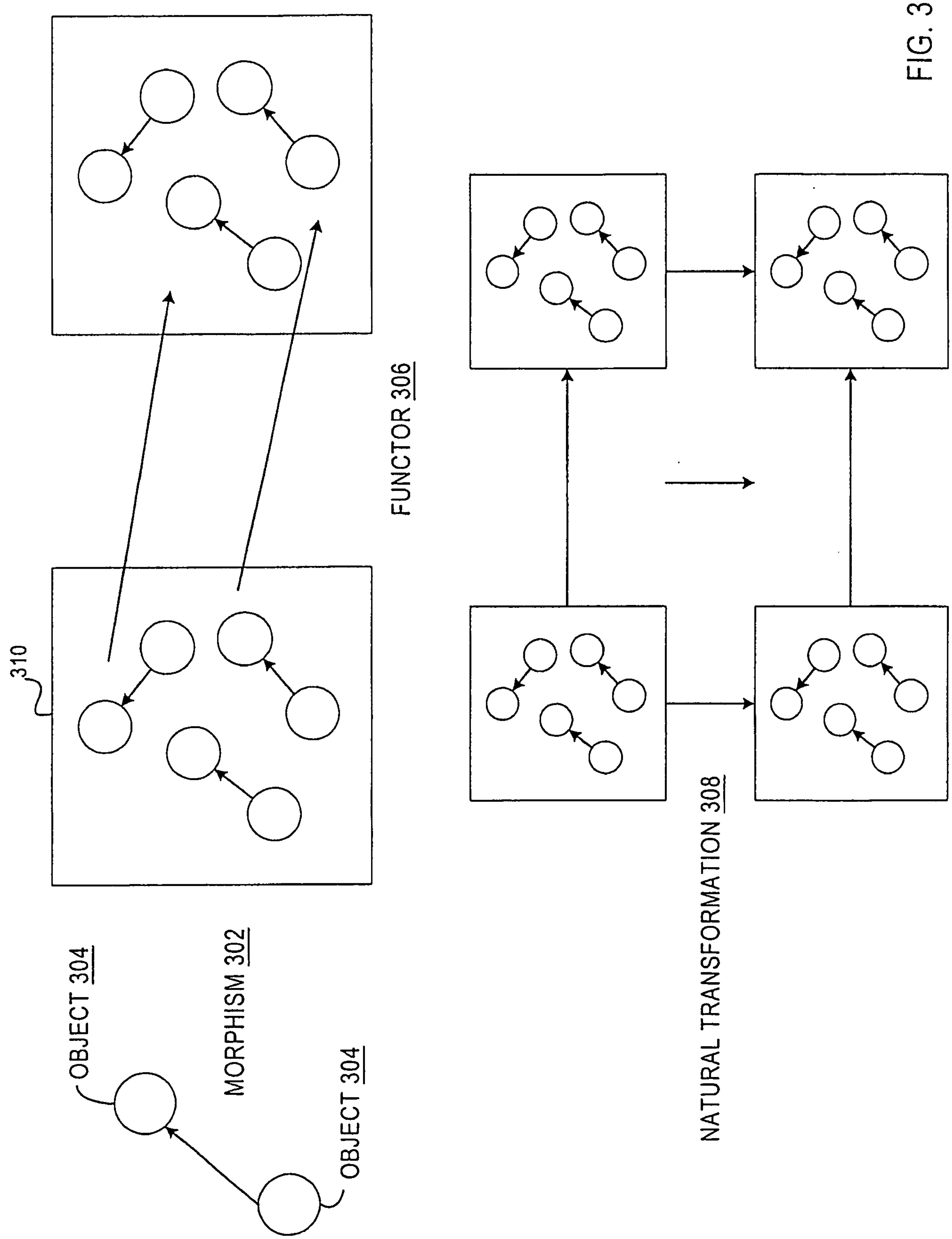
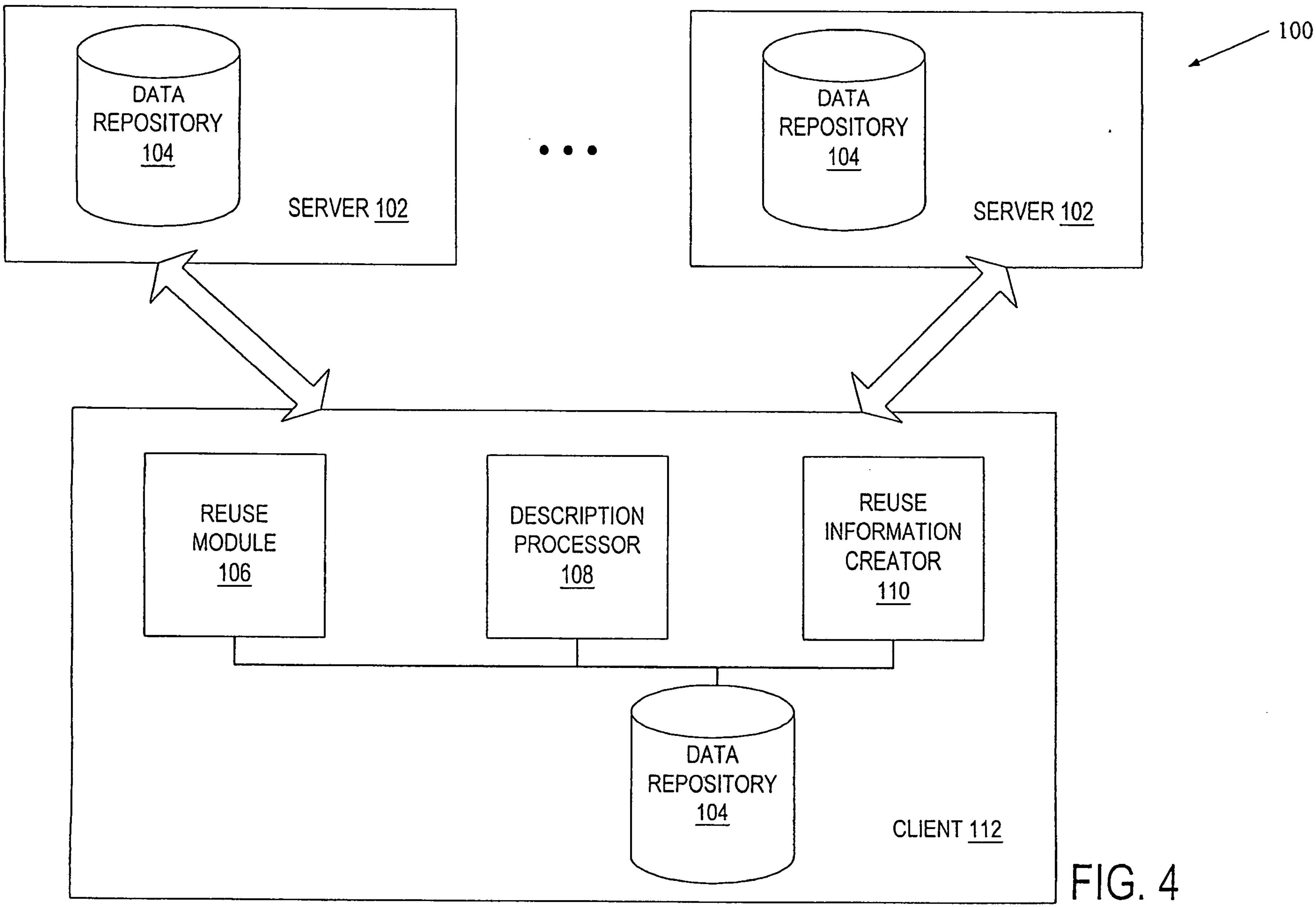

310
FUNCTOR 306
OBJECT 304
MORPHISM 302
OBJECT 304
NATURAL TRANSFORMATION 308

FIG. 3



100
DATA
REPOSITORY
104
SERVER 102
DATA
REPOSITORY
104
SERVER 102
REUSE
MODULE
106
DESCRIPTION
PROCESSOR
108
REUSE
INFORMATION
CREATOR
110
DATA
REPOSITORY
104
CLIENT 112

FIG. 4

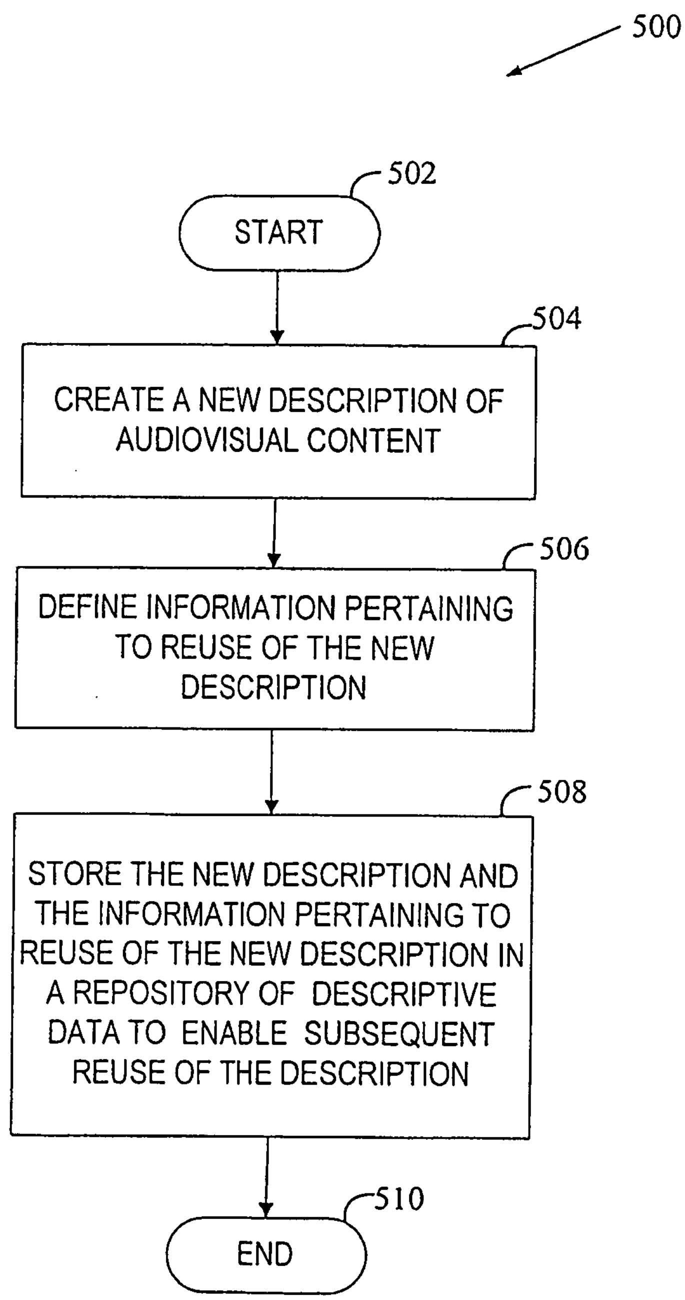

500
502
START
504
CREATE A NEW DESCRIPTION OF AUDIOVISUAL CONTENT
506
DEFINE INFORMATION PERTAINING TO REUSE OF THE NEW DESCRIPTION
508
STORE THE NEW DESCRIPTION AND THE INFORMATION PERTAINING TO REUSE OF THE NEW DESCRIPTION IN A REPOSITORY OF DESCRIPTIVE DATA TO ENABLE SUBSEQUENT REUSE OF THE DESCRIPTION
510
END

FIG. 5

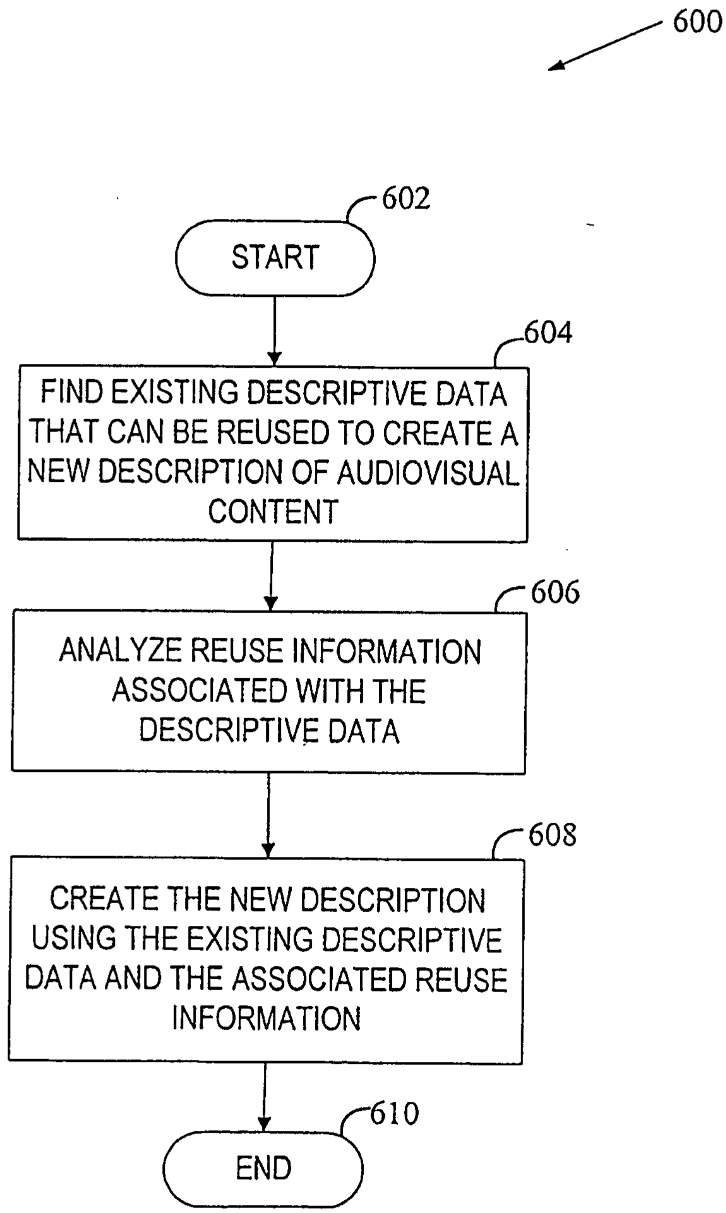

600
602
START
604
FIND EXISTING DESCRIPTIVE DATA THAT CAN BE REUSED TO CREATE A NEW DESCRIPTION OF AUDIOVISUAL CONTENT
606
ANALYZE REUSE INFORMATION ASSOCIATED WITH THE DESCRIPTIVE DATA
608
CREATE THE NEW DESCRIPTION USING THE EXISTING DESCRIPTIVE DATA AND THE ASSOCIATED REUSE INFORMATION
610
END

FIG. 6

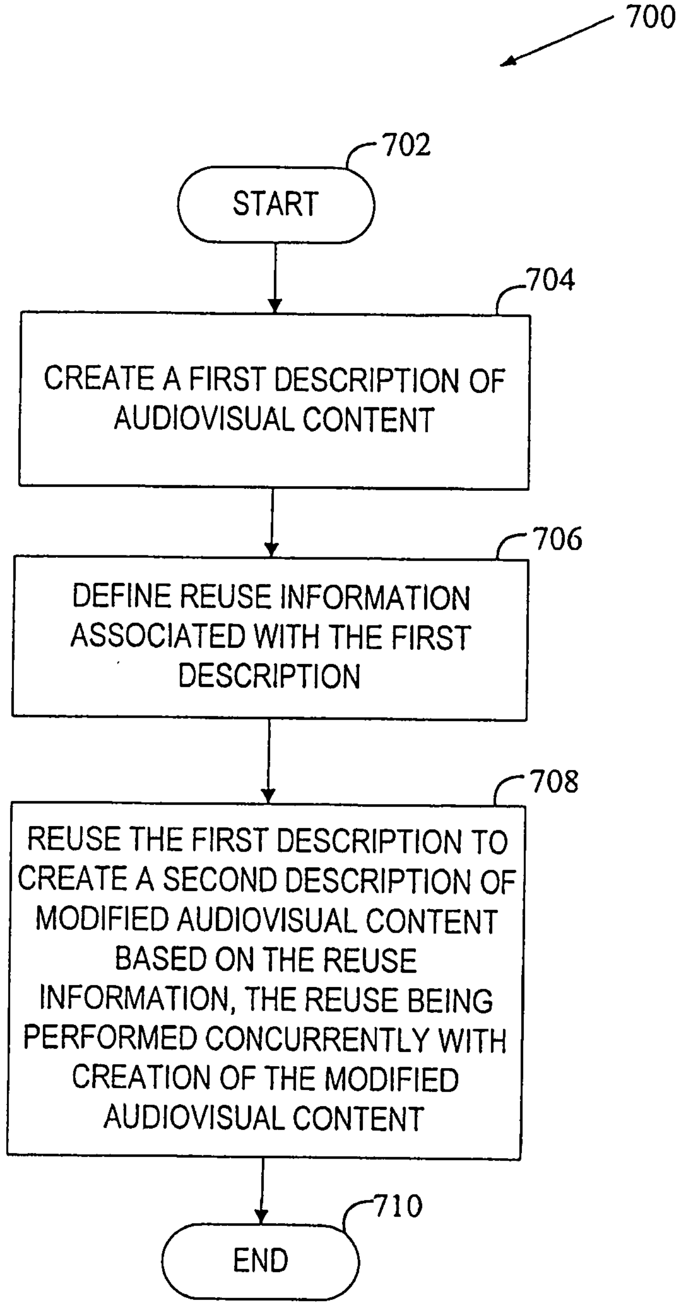

700
702
START
704
CREATE A FIRST DESCRIPTION OF AUDIOVISUAL CONTENT
706
DEFINE REUSE INFORMATION ASSOCIATED WITH THE FIRST DESCRIPTION
708
REUSE THE FIRST DESCRIPTION TO CREATE A SECOND DESCRIPTION OF MODIFIED AUDIOVISUAL CONTENT BASED ON THE REUSE INFORMATION, THE REUSE BEING PERFORMED CONCURRENTLY WITH CREATION OF THE MODIFIED AUDIOVISUAL CONTENT
710
END

FIG. 7

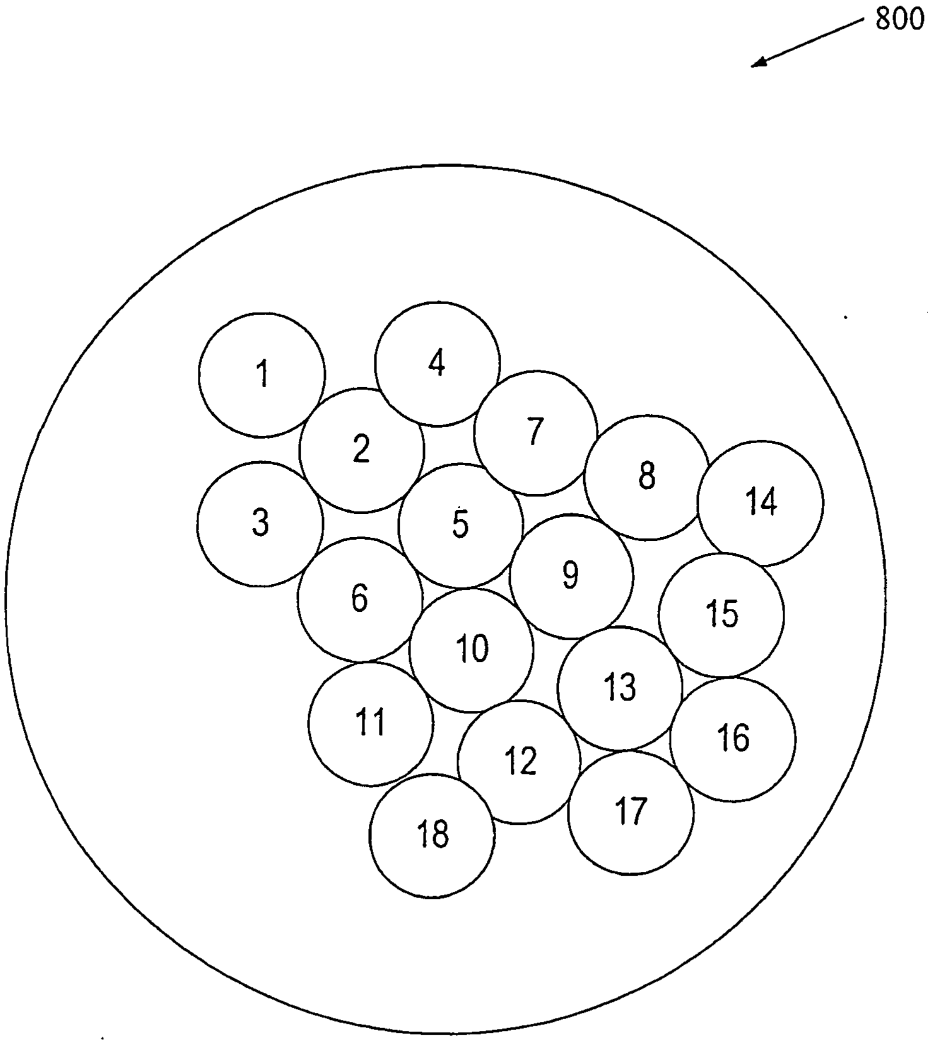


FIG. 8

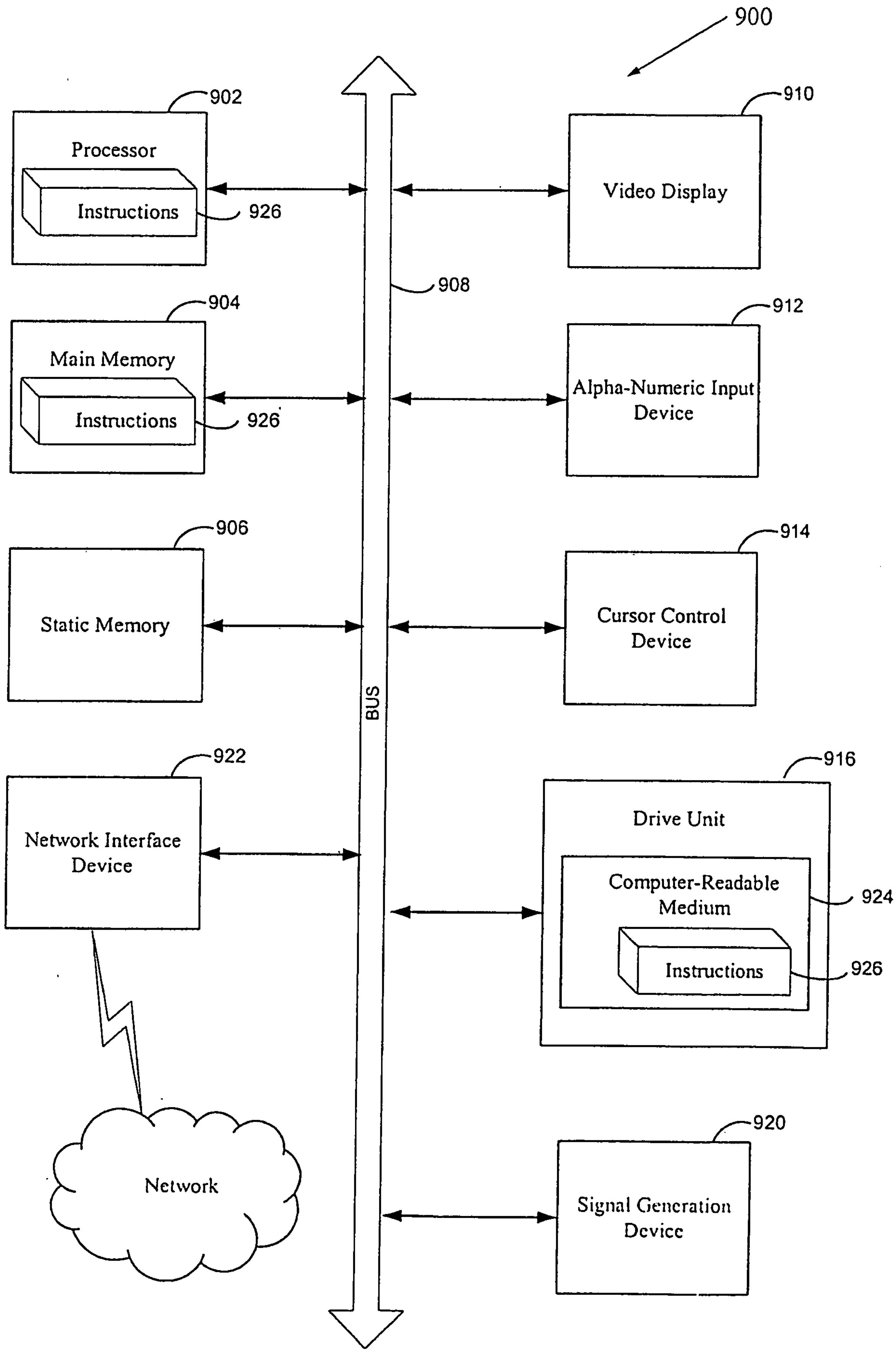

900
902
Processor
Instructions
926
910
Video Display
908
904
Main Memory
Instructions
926
912
Alpha-Numeric Input Device
906
Static Memory
914
Cursor Control Device
BUS
922
Network Interface Device
916
Drive Unit
Computer-Readable Medium
924
Instructions
926
Network
920
Signal Generation Device

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/38395

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : H04N 5/781

US CL : 386/95

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 386/95, 1, 46, 83, 111-112, 125-126; 375/240.01, 240.08; 725/39, 58, 87; 707/102

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 03-125356 A (MIZUGUCHI) 28 May 1991, page 1 of the translation.	1-41
X	JP 07-326089 A (MAENO et al) 12 December 1995, page 1 of the translation.	1-41
X	US 6,070,167 A (QIAN et al) 30 May 2000, Fig. 1 and col. 1, lines 61-67.	1-41

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 13 March 2003 (13.03.2003)	Date of mailing of the international search report 02 APR 2003
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Thai Tran Eugenia Logan Telephone No. (703) 305-4725

Continuation of B. FIELDS SEARCHED Item 3:
EAST
search term: MPEG7, dscription, use, and reuse.